Nguyen Ngoc Linh Chi

LINEAR ALGEBRA – STATISTICS – MACHINE LEARNING cheat-sheet

EQUIVALENT STATEMENT:

- ❋ A is invertible (A has an inverse, is singular)
- ❋ There exists matrix such that $AB = BA = I_n$
- ❋ Transpose of A (A^T) is an invertible matrix
- ❋ $Ax = 0$ has only trivial solution ($x = 0$)
- ❋ The reduced row echelon form of A is I_n
- ❋ $\det(A) \neq 0$
- ❋ The *rows of A* form a basis for $\mathbb{R}^N$
- ❋ The *columns of A* form a basis for $\mathbb{R}^N$
- ❋ A has a full rank, $\operatorname{rank}(A) = n$
- ❋ 0 is NOT eigenvalue of A
- ❋ $Ax = 0$ has a unique solution for each $b \in \mathbb{R}^N$
- ❋ *The null space*$(A) = \{0\}$
- ❋ The *nullity*$(A) = 0$
- ❋ The R/C of A are linearly independent
- ❋ The columns/rows of A span $\mathbb{R}^N$
- ❋ The column/row space of A span $\mathbb{R}^N$
- ❋ The dimension of column/row space if A is n
- ❋ Only vector normal to column/row space = 0

MATRIX MULTIPLICATION:

- ❋ Let $A = (a_{ij})_{m\times p}$ and $B = (b_{ij})_{p\times n}$
- ❋ Pre-multiplication A to B → AB
- ❋ Post-multiplication A to B → BA
- ❋ AB will be $m \times n$ matrix, whose entry (i, j) is: $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ip}b_{pj} = \sum_k^p a_{ik}b_{kj}$.
- ❋ If A; B are diagonal matrices of same size, then AB = BA

INVERSE OF A MATRIX:

- ❋ Scalar product: $(cA)^{-1} = \frac{1}{c}A^{-1}$
- ❋ Inverse; transpose: $(A^T)^{-1} = (A^{-1})^T$
- ❋ Inverse of inverse: $(A^{-1})^{-1} = A$
- ❋ $(AB)^{-1} = B^{-1} \times A^{-1}$
- ❋ $(A_1A_2A_3 \dots A_N)^{-1} = A_N^{-1}A_{N-1}^{-1} \dots A_1^{-1}$
- ❋ $(A^{-1})^N = A^{-N}$
- ❋ If A is invertible; $AC = AB \to C = B$

MATRIX TRANSPOSITION:

- ❋ Let $A = (a_{ij})_{m\times n}$. transpose of A is A^T
- ❋ $(n \times m)$ matrix whose (i, j) entry is a_{ij}
- ❋ A symmetrical matrix if $A = A^T$, $(A^T)^T = A$
- ❋ $(A + B)^T = A^T + B^T$ & $(aA)^T = aA^T$
- ❋ $(AB)^T = B^TA^T$ (NOT A^TB^T)
- ❋ *A is invertible* $\to A^T$ *is invertible*
- ❋ ∀ square matrix, then $\frac{1}{2}(A + A^T)$ is symmetric

DETERMINANT OF SPECIAL MATRICES:

2 × 2: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible → $\det(A) = ad - bc$

Triangular/diagonal matrix

→ det(A) = product of all diagonal entries

Square matrix A → $\det(A) = \det(A^T)$

Square matrix vs 2 same rows or cols $\det = 0$

WAYS TO DETERMINE DETERMINANT:

Elementary Row Operations:

$kR_3 \quad R_1 \leftrightarrow R_3 \quad R_2 + kR_3$

$\det(E_1) = k \quad \det(E_2) = -1 \quad \det(E_3) = 1$

Cofactor expansion:

(i, j) − cofactor of $A = A_{ij} = (-1)^{i+j}\det(M_{ij})$

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 3 & 2 \\ 6 & 5 & 7 \end{vmatrix} = 1(-1)^{1+1}\begin{vmatrix} 3 & 2 \\ 5 & 7 \end{vmatrix} + 2(-1)^{1+2}\begin{vmatrix} 4 & 2 \\ 6 & 7 \end{vmatrix} + 3(-1)^{1+3}\begin{vmatrix} 4 & 3 \\ 6 & 5 \end{vmatrix} = -15 \neq 0$$

PROPERTIES OF DETERMINANT:

$\det(cA) = c^n\det(A)$, $\det(AB) = \det(A)\det(B)$.

$\det(A^{-1}) = \frac{1}{\det(A)}$ & $\det(A) = \det(A^T)$

ADJOINT MATRIX: $\operatorname{adj}(A) = (\text{cofactor of } A)^T$

If A is invertible then $A^{-1} = \frac{1}{\det(A)}\operatorname{adj}(A)$

$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$. cofactor matrix of A:

$A_{11} = \begin{vmatrix} 2 & -1 \\ -1 & 2 \end{vmatrix} \quad A_{12} = -\begin{vmatrix} 0 & -1 \\ 0 & 2 \end{vmatrix} \quad \dots$

Cofactor matrix of A is:

$$\begin{matrix} 3 & 0 & 0 \\ -1 & 4 & 2 \\ -2 & 2 & 4 \end{matrix}$$

The adjoint matrix of A is:

$$\begin{matrix} 3 & -1 & -2 \\ 0 & 4 & 2 \\ 0 & 2 & 4 \end{matrix}$$

CRAMER'S RULE:
$$\begin{matrix} 2 & 1 & 1 & 3 \\ 1 & -1 & -1 & 0 \\ 1 & 2 & 1 & 0 \end{matrix}$$

$$D = \begin{vmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 1 \end{vmatrix}; D_x = \begin{vmatrix} 3 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 2 & 1 \end{vmatrix}$$

$$D_y = \begin{vmatrix} 2 & 3 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{vmatrix}; D_z = \begin{vmatrix} 2 & 1 & 3 \\ 1 & -1 & 0 \\ 1 & 2 & 0 \end{vmatrix}$$

Answer: $x = Dx/D$; $x = Dy/D$; $x = Dz/D$

LU FACTORISATION:

Let A be m x n matrix.

Reduce A to row-echelon form, obtaining U matrix (should be Upper triangular matrix)

$E_KE_{K-1} \dots E_2E_1A = U$

→ $A = E_1^{-1}E_2^{-1} \dots E_{K-1}^{-1}E_K^{-1}U$

$E_1^{-1}E_2^{-1} \dots E_{K-1}^{-1}E_K^{-1} = L$ (lower triangular matrix)

Instead of solving $Ax = B$, solve by $A = LU$

Step 1: solve $Ly = B$ with $y = Ax$

Step 2: solve $Ax = y$

APPLICATION PROBLEM (DEFLECTION):

We know that, by Hooke's Law: $y = Df$

The results are given:

	f_1	f_2	f_3	y_1	y_2	y_3
Ex 1	1	0	1	0.5	0.3	0.5
Ex 2	0	1	2	0.1	0.3	0.7
Ex 3	2	1	0	0.7	0.3	0.1

Question: FIND D

Step 1: $D_{3\times3}f_{3\times3} = y_{3\times3} \to D(u\ v\ w) = (y_1\ y_2\ y_3)$ $(u = (1, 0, 1), v = (0, 1, 2), w = (2,1,0))$

Step 2: we will find a, b, c such that

$au + bv + cw = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. For example:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0.5 & -1 & 0.5 \\ 0 & 1 & 0 & -0.25 & 0.5 & 0.25 \\ 0 & 0 & 1 & 0.25 & 0.5 & -0.25 \end{array}\right)$$

Step 4: combine step 2& 3, we will have:

$$D\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = D\left(\frac{1}{2}u - \frac{1}{4}v + \frac{1}{4}w\right)$$

$$D\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = D\left(-1u + \frac{1}{2}v + \frac{1}{2}w\right)$$

$$D\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = D\left(\frac{1}{2}u + \frac{1}{4}v - \frac{1}{4}w\right)$$

$Du = y_1, Dv = y_2, Dw = y_3$

LINEAR SPAN:

$S = \{u_1, u_2, \dots, u_n\}$ be a set ofvectors in $\mathbb{R}^N$. set containing all linear combination of $\{u_1, u_2, \dots, u_n\}$ is **linear span** of S/linear span of $\{u_1, u_2, \dots, u_n\}$

$\operatorname{span}(S) = \{c_1u_1 + c_2u_2 + \cdots + c_ku_k | c \in \mathbb{R}\}$

To check if $span(S) \neq \mathbb{R}^N$, check whether augmented matrix formed by S; any $x \in \mathbb{R}^N$ is inconsistent (last column is a pivot column)

S_2 *with* $\{u_1, u_2, \dots, u_n\} \in span(S_1)$;

$\{v_1, v_2, \dots, v_n\} \in span(S_2)$

Matrix M: $\boldsymbol{v_1\ v_2\ v_3 \dots v_n | u_1 | u_2 | \dots | u_n}$

$\boldsymbol{Span(S_1) = span(S_2)}$ → prove $Span(S_1) \subseteq span(S_2)$ *AND* $Span(S_2) \subseteq span(S_1)$

$\boldsymbol{Span(S_1) \subseteq span(S_2)}$, then prove that each vector in S_1 linear combination of vectors in S_2

SUBSET; SPAN:

There are 2 main representation of a subset:

Implicit: $V = \{(x, y) | ax + by = 0\}$

Explicit: $V = \{(t, 1 - t) | t \in R\}$

Want to prove that subset = span (S) then we transfer into explicit way to compare

SUBSPACE:

Let V be subset of $\mathbb{R}^N$. If there exists a set of vectors $S = \{u_1, u_2, \dots, u_n\}$ such that $span(S) = V$ then V is said to be a subspace of $\mathbb{R}^N$

V; W is subspaces of $\mathbb{R}^N$

→ $(V + W)$ is subspace of $\mathbb{R}^N$

→ $V \cap W$ is subspace of $\mathbb{R}^N$

→ $V \cup W$ is NOT subspace of $\mathbb{R}^N$

CHECK WHETHER V IS A SUBSPACE OR NOT?

Check V contains vector-0

Check combination of vectors in $V \in V$

ABSTRACT DEFINITION OF SUBSPACE:

Let V be non-empty subset of $\mathbb{R}^N$. Then V is said to be a subspace of $\mathbb{R}^N$ if; only if for ∀ pair vector $(u, v) \in V \to cu + dv \in V$.

LINEAR INDEPENDENCE:

$c_1u_1 + c_2u_2 + \cdots + c_ku_k = 0$. S is called linearly independent set if $c_1 = c_2 = \cdots = c_k = 0$ is ONLY answer.

A is invertible → $Au_1, Au_2, \dots, Au_k$ are linearly independent

REDUNDANCY: $\boldsymbol{u_1, u_2, \dots, u_k}$ are vector taken from $\mathbb{R}^N$. If $\boldsymbol{u_k}$ is a linear combination of $\boldsymbol{u_1, u_2, \dots, u_{k-1}} \to \boldsymbol{u_k}$ is redundant

Check Whether Vector Set Is Independent or Not?

Ex $S = \{u + v, v + w, u + w, u + v + w\}$

$(u, v, w$ are linearly independent $)$

Step 1: we put vector in S in homogenous equation: $c_1(u + v) + c_2(v + w) + c_3(u + w) + c_4(u + v + w) = 0$

Step 2: $(c_1 + c_3 + c_4)u + (c_1 + c_2 + c_4)v + (c_2 + c_3 + c_4)w = 0$

→ This equation only has 1 trivial solution:

→ $c_1 + c_3 + c_4 = c_1 + c_2 + c_4 = c_2 + c_3 + c_4 = 0$ (u, v, w is independent)

NOTE: linear equation has only trivial solution independent. If not → dependent

BASIS: $S \subseteq V$ with $|S| = \dim(V)$.

Let $S = \{u_1, u_2, \dots, u_n\}$ be subset in vector space V. S is called a basis ↔ S is linearly independent, S spans V.

Basis for a vector space $|V|$= smallest possible # vectors that can span V

If U; W: subspaces in $\mathbb{R}^N$, there exists a basis S_1 *for* V & S_2 *for* W such that $S_1 \cap S_2$ is a basis for $V \cap W$, $S_1 \cup S_2$ is a basis for $V + W$

COORDINATE VECTORS:

Let $S = \{u_1, u_2, \dots, u_n\}$ = basis (V) &$v \in V \leftrightarrow$ $v = c_1v_1 + c_2v_2 + \cdots + c_nv_n$

The coefficient $c_1, c_2, \dots, c_k$ Vector $(v_s) = (c_1, c_2 \dots)$ is co-ordinate vector of v relative to basis S.

If S is a basis for v then every vector $v \in V$ has a unique co-ordinate vectors relative to V

V can have different basis → different co-ordinate vectors

EUCLIDIAN VECTORS: $|\boldsymbol{uv}| \leq ||\boldsymbol{u}|| \times ||\boldsymbol{v}||$

$||u + v|| \leq ||u|| + ||v|| \quad \to d(u, w) \leq d(u, v) + d(v, w)$

DIMENSION:

Let V be a vector space having a basis with k vectors.

$|V| > k$ → linearly dependent

$|V| < k$ → cannot span V

The **dimension** of vector space $dim(k)$ = # vectors in basis of v

$Dim(0) = 0$

$W \leq V$ is subspace of V → $\dim(W) \leq \dim(V)$

ROW SPACE; COLUMN SPACE: $A \xrightarrow{ERO} B$

- *Column space* $(A) \neq$ *Column space* (B)
- *Row space*$(A) =$ *row space*(B)
- *Column space* $(A) =$ *row space* (A^T)
- *Row space*$(A) =$ *column space*(A^T)

FIND A BASIS FOR A ROW SPACE:

Find REDUCED row-echelon form is B, then basis of row space of B = basis row space of A

COLUMN SPACE: Let $\boldsymbol{a_1, a_2, \dots, a_k}$ & $\boldsymbol{b_1, b_2, \dots, b_k}$ be columns of A; B respectively. With each $\boldsymbol{Ab_i}$ belongs to column space of A

FIND A BASIS OF COLUMN SPACE OF A:

Find REDCUED row-echelon form of A is B, then choose pivot column of B & take corresponding column(A)

RANK OF A MATRIX:

- ❋ *Rank* $(0) = 0$, *Rank* $(I_n) = n$
- ❋ $m \times n$ matrix, $rank(A) \leq \min(m, n)$
- ❋ $m \times n$ matrix, $rank(A) = \min(m, n)$→ full rank
- ❋ $rank(A) = \dim($*row space of* $A)$
- ❋ $= \dim($*CS of* $A) = rank(A^T)$
- ❋ *Rank* $(AB) \leq \min\{rank(A), rank(B)\}$

COLUMN SPACE; LINEAR SYSTEM:

Theorem: a system of linear equations $Ax = b$ is consistent if; only if b lies in column space of A, or A; augmented matrix have same rank

COLUMN SPACE; LINEAR SYSTEM:

We have equation $Ax = b$:

Consistent → b is linear combination of cols(A) → $b \in CS(A)$

The linear system is **inconsistent**

$\to rank(A|b) = rank(A) + 1$

SPAN-DIMENSION-LINEAR INDEPENDENCE:

If there exists a set $u_1, u_2, \dots, u_p$ that spans V, then $dim(V) \leq p$

If there exists a linearly independent set $u_1, u_2, \dots, u_p$ in V, then $dim(V) \geq p$

If $dim(V) = p$. then there exists a set p + 1 vectors in V that spans V.

NULL-SPACE OF A MATRIX:

Let A be $m \times n$ matrix. $Ax = 0$ be a homogenous linear system

- The solution set of $Ax = 0$ is a subspace of $\mathbb{R}^n$; also solution space of $\mathbb{R}^n$, **null-space (A)**
- $Dim(null - space\ (A)) = \boldsymbol{nullity(A)}$
- $rank(A) + nullity(A) = n$.
- $rank(A)$= # pivot columns of A, nullity (A) = # non-pivot columns of A
- $nullspace\ (A) = nullspace(A^TA)$
- $nullity(A) = nulity\ (A^TA)$
- $nullity(A) \neq nullity(A^TA)$
- $rank(A) = rank(A^TA)$
- $rank(A) = rank(AA^T)$

RANK VS. NULLITY:

$(Rank(A) = \#pivot\ columns)$

$= \#leading\ entries\ in\ R$

$Nullity(A) = \#non-pivot\ cols = \#arbitrary\ parameters\ in\ Ax = 0$

SOLUTION FOR Ax = b WITH NULL-SPACE:

Let x_h be general solution for $Ax = 0$, let x_p be solution to equation: $Ax = b$, general solution to $Ax = b$ is $x_h + x_p$

Orthogonality:

$S = \{u_1, \dots, u_n\}$ orthogonal basis, $\forall\ w \in V$:

$$w = \frac{w.u_1}{||u_1||^2}.u_1 + \frac{w.u_2}{||u_2||^2}.u_2 + \cdots$$

$$\text{coordinate vector}(w_s) = \left(\frac{w.u_1}{||u_1||^2}, \frac{w.u_2}{||u_2||^2}, \dots\right)$$

ORTHOGONAL PROJECTION:

Let V be a subspace of $\mathbb{R}^N$; $w \in \mathbb{R}^N$ then if $\{u_1, u_2, \dots, u_n\}$ is an orthogonal basis for V, then projection of w onto V is p:

$$p = \frac{w.u_1}{||u_1||^2}.u_1 + \frac{w.u_2}{||u_2||^2}.u_2 + \cdots$$

FIND DISTANCE FROM A POINT TO A LINE/PLANE:

For a line, we only have 1 basis:

Step 1: projection $p = \frac{w.u}{||u||^2}.u$

Step 2: compute distance $d = ||w - p||$

For a plane, we have 2 bases, however, alternative method: we can use "already known" orthogonal vector:

$$ax + by + cz = d, \vec{n} = (a, b, c)$$

Step 1: projection w onto n, $p = \frac{w.n}{||n||^2}.n$

Step 2: take length of p: $d = ||p||$

Orthogonal/orthonormal:

ORTHOGONAL/ORTHONORMAL SET:

- A set of vectors $\{\vec{v_1}, \dots\}$ are mutually **orthogonal** is every pair of vectors is orthogonal $\vec{v_i} \times \vec{v_j} = 0, \forall i \neq j$
- Orthogonal/orthonormal set is a basis for $\mathbb{R}^N$ because:

They are set of non-zero vectors

Linearly independent set

$|orthogonal/orthonormal\ set| = N$

MATRIX WITH ORTHONORMAL COLS:

$A \in \mathbb{R}^{m\times n}$ (A **don't** have to be square matrix) has orthonormal vectors columns if its Gram matrix is I

Orthogonal matrices:

$A \in R^{n\times n}$ has orthonormal columns $(m = n)$

1. Properties: Q is orthogonal matrix

- Q^{-1} is also orthogonal matrix
- $\det(Q) = \pm1$
- If λ is an eigenvalue of Q, then $|\lambda| = 1$
- *If* Q_1 & Q_2 are $n \times n$ orthogonal matrix, Q_1Q_2 is also orthogonal matrix
- **Rows of A** are an orthogonal matrix

NOTE: $A \in R^{m\times n}$ has orthonormal cols $(m > n)$ is NOT **orthogonal matrix**

$$(A^TA = I\ AND\ AA^T \neq I)$$

Product of orthogonal matrices:

If $A_1, A_2, A_3, \dots, A_k$ are **orthogonal matrices**; of equal size, then product: $A = A_1A_2 \dots A_k$ is **orthogonal matrix**

Linear equation with orthogonal matrix:

$$Ax = b \to x = A^{-1}b = A^Tb$$

BEST APPROXIMATION:

p is best approximation of $u \in V$ if

$d(u, p) \leq d(u, v)\ \forall v \in V$: p is projection

ORTHOGONAL basis (GRAM-SCHMIDT process):

For a line, span{u}

For plane, $u_1 = v_1$ & $v_2 = \frac{u_2.v_1}{||v_1||^2}.v_1$

For 3D space,

$$v_3 = u_3 - \left(\frac{u_3.v_1}{||v_1||^2}.v_1 + \frac{u_3.v_2}{||v_2||^2}.v_2\right)$$

$$(v_1, v_2\ like\ in\ the\ plane)$$

NOTE: if you are asked to find **orthonormal** then have to make each vector in **orthogonal** become **unit vector**

MATRIX-VECTOR PRODUCT: $f(x) = Ax$

$(Ax)^T(Ay) = x^TA^TAy = x^Ty$

$||Ax|| = \left((Ax)^T(Ax)\right)^{\frac{1}{2}} = (x^Tx)^{\frac{1}{2}} = ||x||$

$||Ax - Ay|| = ||x - y||$

$$\angle(Ax, Ay) = \cos^{-1}\left(\frac{(Ax)^T(Ay)}{||Ax||||Ay||}\right) = \angle(x, y)$$

$Q\vec{x}Q\vec{y} = \vec{x}\vec{y}$

$$\begin{bmatrix} a_1^Ta_1 & a_1^Ta_2 & \cdots & a_1^Ta_n \\ a_2^Ta_1 & a_2^Ta_2 & \cdots & a_2^Ta_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n^Ta_1 & a_n^Ta_2 & \cdots & a_n^Ta_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_n$$

$A^TA = I$ (left inverse), $m > n \to AA^T \neq I$

A has a linearly independents columns:

$$Ax = 0 \to A^TAx = x = 0$$

ORTHONORMAL VECTORS:

- The vectors have unit norm $||a_i|| = 1$
- Mutually orthogonal: $a_i^T \times a_j = 0$
- **For ex,** a set of orthogonal $\{u_1, u_2, u_3, u_4\}$ → span $\{u_4\}$ = ∀ vectors orthogonal to span $\{u_1, u_2, u_3\}$

ORTHOGONAL complements & ORTHOGONAL projections

Let W be a subspace of Rn. **set of all vectors** that are **orthogonal to W** is called **orthogonal complement of W,** denoted $W^\perp$. That is, $W^\perp = \{\vec{v} \in \mathbb{R}^n | \vec{v}.\vec{w} = 0\ \forall\ \vec{w} \in W\}$

Properties: Let W be a subspace of $\mathbb{R}^n$

- $W^\perp$ is a subspace of $\mathbb{R}^n$
- $(W^\perp)^\perp = W$ & $\dim(W) + \dim(W^\perp) = n$
- $W^\perp \cap W = \{\vec{0}\}$ & W = span $\{w_1 \dots, w_n\}, \vec{v} \in W^\perp \leftrightarrow \vec{v}.\vec{w_i} = 0\ \forall i$
- Let A: $m \times n$ matrix. Then orthogonal complement of row space of A is null space of A; orthogonal complement of column space of A is null space of A^T: $(row(A))^\perp = null(A)$ & $(col(A))^\perp = null(A^T)$

LEAST SQUARE SOLUTION:

- $Ax = b$ is **inconsistent** → b is not in column space of A → project b onto column space of A → solve $Ax = p$ (projection of b onto A)
- *A vector* $u \in \mathbb{R}^N$ is least square *solution*: $||b - Au|| \leq ||b - Av||\ \forall\ v \in \mathbb{R}^n$
- **Theorem:** let $Ax = b$ be a linear system. Then x is a least square solution to $Ax = b$ if; only if x is a solution to $\boldsymbol{A^TAx = A^Tb}$
- If linear system is **consistent** then solution set of $Ax = b$ is **same** as $A^TAx = A^Tb$

QR FACTORIZATION:

Step 1: We have A as a $m \times n$ matrix, treat each column as a vector $A = (u_1, u_2, u_3)$

Step 2: find **orthonormal basis** (w_1, w_2, w_3) (unit vector of orthogonal basis)

Step 3: Write each of (u_1, u_2, u_3) as a linear combination of (w_1, w_2, w_3). They should be in form like this:

$u_1 = a_1 \times w_1, u_2 = b_1 \times w_1 + b_2 \times w_2$.

$u_1 = c_1 \times w_1 + c_2 \times w_2 + c_3 \times w_3$.

Step 4: write A = QR (Q is orthogonal matrix), R is triangular matrix, **+ entries**

- $A = (u_1, u_2, u_3) = (w_1, w_2, w_3)\begin{pmatrix} a_1 & b_1 & c_1 \\ 0 & b_2 & c_2 \\ 0 & 0 & c_3 \end{pmatrix}$

QR factorisation + LEAST SQUARE solution:

NOTE: Q: orthogonal matrix → $Q^TQ = I$

We have $A = QR$ → compute solution for Least Square solution: $A^TAx = A^Tb$, $x = R^{-1}Q^Tb$

EIGENVALUE & EIGENVECTOR:

- $\boldsymbol{Au = \lambda u}$, λ is called **eigenvalue**; u is **eigenvector** associated with **eigenvalue $\boldsymbol{\lambda}$**
- The **eigenvalue** of a **diagonal matrix** are all entries on main diagonal
- λ is an **eignvalue of $\boldsymbol{A}$** *then*:

λ^n is an eigenvalue of $A^n \to A^nu = \lambda^nu$

$1/\lambda$ is eigenvalue of A^{-1} $(A\ is\ invertible)$

- $n \times n$ matrix has eigenvalues $\lambda_1, \lambda_2, \dots$, eigenvectors $\vec{v_1}, \vec{v_2}, \dots$ If $\vec{x} \in \mathbb{R}^n$; $\vec{x} = c_1\vec{v_1} + c_2\vec{v_2} + \cdots$, $\forall k \in \mathbb{Z}$: $A^k\vec{x} = c_1\lambda_1^k\vec{v_1} + c_2\lambda_2^k\vec{v_2} + \cdots$

FIND AN EIGENVALUE OF A MATRIX:

- Eigenvalue of A: λ that makes matrix $(\lambda I - A)$ singular → $\det(\lambda I - A) = 0$
- $\det(\lambda I - A)$: **Characteristic polynomial** A
- $det(\lambda I - A) = 0$: **Characteristic equation**

EIGENSPACE:

- $(\lambda I - A)x = 0$: singular → system has infinite solution.
- The solution space $(\lambda I - A)x = 0$: **eigen-space** of A associated with eigenvalue $\lambda\ (E_\lambda)$ → E_λ = ALL eigenvectors of A
- E_λ is called a null-space of $(\lambda I - A)$
- For each λ, we find $E_\lambda = span\{u_1, u_2, \dots\}$

Diagonalisation:

- P *diagonalise* $A \leftrightarrow \exists\ P |\ P^{-1}AP = D$, D is a diagonal matrix with entries are $\lambda_1, \lambda_2, \dots$
- $n \times n$ *matrix* A is diagonalizable if; only if A has n linearly independent eigenvectors /n distinct eigenvalues

CHECK WHETHER A IS DIAGONALIZABLE?

Step 1: Find eigenvalues $\det(\lambda I - A) = 0$

- $det(\lambda I - A) = (\lambda - \lambda_1)^{r_1} \dots (\lambda - \lambda_i)^{r_i} \dots$
- $r_1 + r_2 + \cdots + r_n = n$ (order of matrix A)

Step 2: For each λ_i, we find a basis for eigenspace E_λ

- r_i: **algebraic multiplicity** of eigenvalue λ_i
- $\dim(E_\lambda)$: **geometric multiplicity** of λ_i
- Then for each λ_i, $\dim(E_\lambda) \leq r_i$

Step 3: Let $S = S_1 \cup S_2 \cup S_3 \dots \cup S_k$, $|S|$ is number of vectors in S.

- If $|S| < n$ → A is NOT diagonalizable
- If $|S| = n$ → A is diagonalizable.

Equality happens: $|S_{\lambda i}| = \dim(E_{\lambda i}) = r_i$

$\leftrightarrow |S| = |S_1 \cup \dots \cup S_k| = r_1 + r_2 + \cdots + r_n = n$

↔ **geometric multiplicity = algebraic multiplicity** for all eigenvalues of A

- If $|S| > n$ → NEVER HAPPEN

ORTHOGONAL *diagonalise* SYMMETRIC matrix:

- A square matrix A is orthogonally diagonalizable ↔ ∃ orthogonal matrix Q such that $Q^TAQ = D$ is a diagonal matrix
- **If a matrix A is orthogonally diagonalizable, then A is symmetric** (NOTE: Q is orthogonal matrix).

→ $Q^TAQ = D$; → $A = QDQ^T (since\ Q^T = Q^{-1})$ → $A^T = (QDQ^T)^T = (Q^T)^T D^T Q^T = QDQ^T = A, so$ A is symmetric.

Eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal.

METHOD for ORTHOGONAL diagonalization of a symmetric matrix.

1. Find eigenvalues of A.
2. Find eigenspace for each eigenvalue.
3. For repeated eigenvalues (when dimension of eigenspace > 1

→ apply Gram–Schmidt orthogonalization to find an orthogonal basis.

4. These orthogonal bases of eigenspaces form an orthogonal basis of $\mathbb{R}^n$.
5. Normalize, dividing each vector of basis by its length.
6. We have: $Q^TAQ = Q^{-1}AQ = D$, where D is diagonal with eigenvalues of A

REAL LIFE PROBLEM:

- $X_n = AX_{n-1} = A^{n-1}.X_1, A\ is\ a\ matrix$
- $A = PDP^{-1}\ with\ D = P^{-1}AP$
- D is a diagonal matrix; diagonal entries of D is $\lambda_1, \lambda_2, \ldots$

Then we have: $A^n = PD^nP^{-1}$. D is diagonal matrix so $D^n = (all\ diagonal\ entries)^n$

(APPLICATION)TILES:

- I have 3 kinds of tiles: 1×1red-colored tiles (1R), 1×2 blue-colored tiles (2B); 1×2 green-colored tiles (2G).
- Let b_n = # different ways to tile a $1 \times$ n pavement.

For example,

- $b_1 = 1$; $b_2 = 3$

Question: Find b_n.

ANSWER

Step 1: find relation between b_{n+1}, b_n, b_{n-1}

We have already known b_n ways to tile $1 \times$ n pavement. Now if we want to tile $1 \times (n+1)$ then: we can add on 1R, or take out 1R from b_n make it b_{n-1}; add 2G/2B. since they are 3 ways (1 for b_n; 2 for b_{n-1}) then total ways is: $b_{n+1} = b_n + 2b_{n-1}$

Step 2: set up $X_n = AX_{n-1} = A^{n-1}X_1$

$X_n = \begin{pmatrix} b_n \\ b_{n+1} \end{pmatrix} = \begin{pmatrix} b_n \\ b_n + 2b_{n-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} b_{n-1} \\ b_n \end{pmatrix} = A.X_{n-1}$.

Step 3: set up A^n:

Find eigenvalue: $\lambda(\lambda - 1) - 2 = 0$

P? $E_{\lambda 1} = span\left\{\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right\}, E_{\lambda 2} = span\left\{\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right\}$

$P = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}$.

Find $D = P^{-1}AP$ should be $\begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$

Set up $A^n = PD^nP^{-1}$

$X_n = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} 2^n & 0 \\ 0 & (-1)^n \end{pmatrix}\begin{pmatrix} 1/3 & 1/3 \\ 2/3 & -1/3 \end{pmatrix}$

$= \begin{pmatrix} 2/3 \times 2^n + 1/3 \times (-1)^n \\ 2/3 \times 2^{n+1} + 1/3 \times (-1)^{n+1} \end{pmatrix}$.

Alternative method:

$X_n = \begin{pmatrix} s2^n + t(-1)^n \\ u2^{n+1} + v(-1)^{n+1} \end{pmatrix}$

When $n = 1$:

$s = u, t = v \rightarrow X_1 = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 2s - 1t \\ 4s + 1t \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

$\rightarrow s = 2/3\ \&\ t = 1/3$

Problem 1: rate of change:

$$\mathbf{rate = rate\ in - rate\ out}$$

CONIC EQUATION:

We have $Ax^2 + Bxy + Cy^2 = k \in \mathbb{R}$. general form of conic section:

$$(x - a \quad y - b)\begin{pmatrix} A & B/2 \\ B/2 & C \end{pmatrix}\begin{pmatrix} x - a \\ y - b \end{pmatrix} = 0$$

LINEAR DIFFERENTIAL EQUATION:

- Differential equation: $X' = AX$
- If λ is an egienvalue of A, associated with x_1 is vectorspace then $X_1 = e^{\lambda_1}x_1$ → general solution is: $k_1e^{\lambda_1}x_1 + k_2e^{\lambda_2}x_2$ (x_1, x_2) are vectors. Then we use initial condition to find (k_1, k_2)

FUNDAMENTAL SET: *Square matrix A:*

- ∃(fundamental solution set) $Y' = AY$
- n linearly independent functions in fundamentall set S
- S is an n-dimensional vector space of functions
- If vector Y_0 is specified, initial value problem is to construct a unique Y such that $Y' = AY$ & $Y(0) = Y_0$

PROPERTIES OF COMPLEX VECTORS:

- $\overline{ku} = \bar{k}\,\bar{u}$ & $\overline{u + v} = \bar{u} + \bar{v}$
- $\overline{u - v} = \bar{u} - \bar{v}$
- $u.v = u_1.\overline{v_1} + u_2.\overline{v_2} + \cdots + u_n.\overline{v_n}$
- $||u|| = \sqrt{v.v} = \sqrt{|v_1|^2 + |v_2|^2 + \cdots |v_n|^2}$
- $u.v = \overline{v.u}$ (asymmetric property)
- $u.kv = \bar{k}(u.v)$

REVIEW ABOUT COMPLEX NUMBER:

The polar form of a complex number:

$z = r(cos\theta + isin\theta) = a + bi = re^{i\theta}$

Complex exponential form:

$e^{a+ib} = e^a \times e^{ib} = e^a(\cos b + i\sin b)$

COMPLEX EIGENVALUE: $\boldsymbol{\lambda = a + ib}$

- If λ is an eigenvalue of A; x is eigenvector associated with λ, then $\bar{\lambda}$ is an eigenvalue of A; $\bar{x}$ is eigenvector associated with $\bar{\lambda}$
- Furthermore, we all know that $e^{\lambda t}x\ and\ e^{\bar{\lambda}t}\bar{x}$ are both conjugate solutions of $Y' = AY \rightarrow$ linear combination of $e^{\lambda t}x\ and\ e^{\bar{\lambda}t}\bar{x}$ = a solution to this equation (if we don't have initial condition)
- Consider following linear combination of $e^{\lambda t}x\ and\ e^{\bar{\lambda}t}\bar{x}$:

$$Y_1 = \frac{1}{2}\left(e^{\lambda t}x + e^{\bar{\lambda}t}\bar{x}\right) = Re\left(e^{\lambda t}x\right) \in R$$

$$Y_2 = \frac{1}{2}\left(e^{\lambda t}x - e^{\bar{\lambda}t}\bar{x}\right) = Im\left(e^{\lambda t}x\right) \in R$$

$$Y_1 + Y_2 = e^{\lambda t}x = e^{(a+ib)t}x$$

$$= e^{at}(\cos bt + i\sin bt)(\mathrm{Re(x)} + i\mathrm{Im(x)})$$

Application (Complex Eigenvalue):

$\lambda = 2 \pm 5i. Y_0 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, E_\lambda = span\left\{\begin{pmatrix} i \\ 2 \end{pmatrix}\right\}$.

$Let\ x = \begin{pmatrix} i \\ 2 \end{pmatrix} \rightarrow e^{\lambda t}x = e^{(2+5i)t}x$

$= e^{2t}(\cos 5t + i\sin 5t)\begin{pmatrix} i \\ 2 \end{pmatrix}$

$= \begin{pmatrix} -e^{2t}\sin 5t \\ 2e^{2t}\cos 5t \end{pmatrix} + i\begin{pmatrix} e^{2t}\cos t \\ 2e^{2t}\sin 5t \end{pmatrix}$

$Y_1 = Re\left(e^{\lambda t}x\right) = \begin{pmatrix} -e^{2t}\sin 5t \\ 2e^{2t}\cos 5t \end{pmatrix}\ and$

$Y_2 = Im\left(e^{\lambda t}x\right) = \begin{pmatrix} e^{2t}\cos t \\ 2e^{2t}\sin 5t \end{pmatrix}$.

The general solution is $c_1Y_1 + c_2Y_2$. When t = 0 then:

$$\mathbf{c_1}\begin{pmatrix} \mathbf{0} \\ \mathbf{2} \end{pmatrix} + \mathbf{c_2}\begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{3} \\ \mathbf{3} \end{pmatrix} \rightarrow \mathbf{c_1}\ \&\ \mathbf{c_2}$$

PROBLEM 2: $\mathbf{x'' + ax' + bx = 0}$:

Step 1: Let $y = x'\ and\ z = x$.

Step 2: We have: $y' + ay + bz = 0$.

Step 3: Find another equation relating to z', y, z, we have $z' = y$

$\begin{pmatrix} y' \\ z' \end{pmatrix} = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y \\ z \end{pmatrix} \rightarrow A = \begin{pmatrix} -a & -b \\ 1 & 0 \end{pmatrix}$

Then solve as usual.

EXERCIES:

IDEMPOTENT MATRICES:

A matrix A is said to be idempotent when: $A^2 = A$

A idempotent ↔ $(I - A)$ is idempotent

A is idempotent → $(I + A)$ is invertible

SUBSPACES:

Let V; W: subspaces of $\mathbb{R}^n$. Define:

$$V + W = \{v + w | v \in V\ and\ w \in W\}\ (1)$$

Show that $V + W$ is a subspace of $\mathbb{R}^n$

$V = span\{v_1, \ldots\}, W = span\{w_1, \ldots\}$

$V + W = \{v + w | v \in V\ and\ w \in W\}$

$= \{a_1v_1 + \cdots a_mv_m + b_1w_1 + \cdots + b_nw_n$

$|a_1, \ldots, a_m, w_1, \ldots, w_n\} \rightarrow (1)$

Show that $V \cap W$ is a subspace of $\mathbb{R}^n$

Use abstract definition of subspace

Step 1: Show that $V \cap W$ non empty, true because 0 always belong in $V \cap W$

Step 2: Let u; v be any two vectors in $V \cap W$; let a; b be any real numbers.

u; v ∈ V, $au + bv$ ∈ V. Similarly, $au + bv$ ∈ W. Thus $au + bv \in V \cap W$.

By abstract definition of subspaces, $V \cap W$ is a subspace of $\mathbb{R}^n$

If V; W: subspaces, ∃ basis S_1 for V; basis S_2 for W such that $S_1 \cup S_2$ is basis $V + W$ & $S_1 \cap S_2$ is basis $V \cap W$

$\{u_1, u_2, \ldots, u_k\}$ be a basis for $V \cap W$. By adding in vectors successively, there exists vectors $\{v_1, v_2, \ldots, v_m\} \in V | \{u_1, \ldots u_k, v_1, \ldots, v_m\}$ is a basis for V; there exists vectors $\{w_1, w_2, \ldots, w_n\} \in V | \{u_1, \ldots u_n, w_1, \ldots, w_n\}$ is a basis for V.

V + W = span $\{u_1, \ldots u_n, v_1, \ldots, v_m, w_1, \ldots, w_n\}$

Consider vector equation

$$\sum_{i=0}^{k} a_iu_i + \sum_{i=0}^{m} b_iv_i + \sum_{i=0}^{n} c_iw_i = 0(1)$$

$$\rightarrow \sum_{i=0}^{k} a_iu_i = -\left(\sum_{i=0}^{m} b_iv_i + \sum_{i=0}^{n} c_iw_i\right) \in V \cap W$$

$$\rightarrow \sum_{i=0}^{n} c_iw_i = -\sum_{i=0}^{k} d_iu_i \rightarrow c_i = d_j$$

$$(1) \rightarrow \sum_{i=0}^{k} a_iu_i + \sum_{i=0}^{m} b_iv_i = 0$$

$\rightarrow \{u_1, u_2, \ldots, u_k, v_1, v_2, \ldots, v_m, w_1, w_2, \ldots, w_n\}$ are linearly independent, basis of $V + W$

Let $x_1, x_2, \ldots$ independent vectors in $\mathbb{R}^n$.

If A is invertible matrix $\rightarrow Ax_1, Ax_2, \ldots, Ax_n$ are linearly independent (1)

$c_1Ax_1 + c_2Ax_2 + \cdots + c_nAx_n = 0$

$\rightarrow A(c_1x_1 + c_2x_2 + \cdots + c_nx_n) = 0$

$\rightarrow A^{-1}A(c_1x_1 + c_2x_2 + \cdots + c_nx_n) = 0$

$\rightarrow c_1x_1 + c_2x_2 + \cdots + c_nx_n = 0$

$\rightarrow c_1 = c_2 = \cdots = c_n\ uniquely = 0 \rightarrow (1)$

RANK, NULLITY:

$a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_p$ be cols of A; B. Show: $Ab_i \in Column\ Space(A)$

$$Ab_i = (a_1 \ldots a_n)\begin{pmatrix} b_{i1} \\ \vdots \\ b_{im} \end{pmatrix} = b_{i1}a_1 + b_{i2}a_2 + \cdots$$

$$\rightarrow AB = Ab_i\ is\ the\ column\ space\ (A)$$

Show that $rank(AB) \leq rank(A)$

$Ab_i is\ the\ column\ of\ AB, Ab_i \subseteq CS(A)$

$C = span\{Ab_1 ..\} \subseteq D = span\{a_1 .. a_n\}$

$\dim(span\{Ab_1 ..\}) \leq \dim(span\{a_1 ..\})$

$\dim(CS(AB)) \leq \dim(CS(A))$

$rank(AB) \leq rank(A)$

Show that $rank(AB) \leq rank(B)$

$rank(AB) = rank(AB)^T = rank(B^TA^T)$

$\leq rank(B^T) = rank(B)$

$rank(A) + rank(B) - n \leq rank(AB) \leq \min\{rank(A); rank(B)\}$

NULLSPACE:

Show $nullspace$ of A = $nullspace\ A^TA$

Let u: vector of $nullspace$ of A, $Au = 0 \rightarrow A^TAu = 0$

Let v be vector of $nullspace\ A^TA$

$Av = (b_1\ b_2 \ldots b_m)^T$; $A^TAv = 0$

$\rightarrow (A^Tv)(Av) = v^TA^TAv = 0$

$\rightarrow b_1^2 + b_2^2 + \cdots + b_m^2 = 0 \rightarrow Av = 0$

→ nullspace of A^TA is subspace of $nullspace$ of A

Explain why every vector in $nullspace$ of B is also in $nullspace$ of AB. Is this also true for every vector in $nullspace$ of A

Suppose a vector x is in $nullspace$ of B, then we get $Bx = 0$. By matrix multiplication, $ABx = 0 \rightarrow$ x is also in nullspace of AB.

This is NOT true for every vector in $nullspace$ of A. Take $A = [1,0; 0,0], B = [1,1; 0,0]$; $x = (0,1)$. Then $Ax = 0$, but x is not in nullspace of AB.

Let A; B be **n × n** *matricies.*

a) Show that AB = 0 if; only if column space of B is a subspace of $nullspace$ of A

For this part of problem let columns of B be $b_1 .. b_n$. Then $B = [b_1 .. b_n]$. $AB = 0 \Leftrightarrow A[b_1 .. b_n] = [Ab_1 .. Ab_n] = 0 \Leftrightarrow Ab_i = 0 \Leftrightarrow$ *All elements of column space of B must be contained in nullspace of A.*

b) Show that if AB = 0, then sum of ranks of A; B cannot exceed n.

$AB = 0 \rightarrow CS(B)$ is a subspace of nullspace of A.

$rank(B) = dim(CS(B))$

$\rightarrow dim(Null(A)) \geq dim(rank(B))$.

$\rightarrow dim(Null(A)) + dim(rank(A)) = n\ and$ $dim(Null(B)) + dim(rank(B)) = n$.

$\rightarrow n - dim(rank(A)) \geq dim(rank(B))$.

$\rightarrow n \geq dim(rank(A) + dim(rank(B))$

Let A be an $m \times m$ non-singular matrix; B be an $m \times n$ matrix. Prove that AB; B have same null space.

Step 1: Null space(B) is a subset of null space (AB).

If v is in null space of B, then $Bv = 0$; hence, $ABv = 0$. Thus, v is also in null space of AB.

Step 2: Null space(AB) contained in null space(B)

If v is in null space of AB → ABv = 0. matrix A is non-singular; $A(Bv) = 0$. It follows that $Bv = 0$; hence v is also in null space of B

$$\rightarrow \dim(CS(AB)) = \dim(CS(B))$$

EIGNEVALUES; IDENTITY MATRIX

A is a diagonalizable $n \times n$ matrix; has only 1; −1 as eigenvalues. *Show that $A^2 = I_n$.*

Since A diagonalizable has ± 1 eignevals $A = PDP^{-1}$

$\rightarrow A^2 = (PDP^{-1})(PDP^{-1}) = PD^2P^{-1}$

$(D^2 = I_2) \rightarrow A^2 = I_n$

THE EXPONENTIAL OF A MATRIX:

$$e^A = I + A + \frac{1}{2!}A^2 + \cdots = \sum_0^{\infty} \frac{1}{n!}A^n$$

Compute e^A

Step 1: $Diagonalise$ matrix

Step 2: $A^n = P\begin{pmatrix} 2^n & 0 \\ 0 & 4^n \end{pmatrix}P^{-1}$

$$= \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2^n & 0 \\ 0 & 4^n \end{pmatrix}\begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

Step 3: convert into exponential form:

$$e^A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} e^2 & 0 \\ 0 & e^4 \end{pmatrix}\begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

Step 4: using matrix multiplication to calculate:

$$e^A = \frac{1}{2}\begin{pmatrix} e^4 + e^2 & e^4 - e^2 \\ e^4 - e^2 & e^4 + e^2 \end{pmatrix}$$

RECURRENCE OF DETERMINANT:

$$\begin{pmatrix} 3 & 1 & & & & & \\ 1 & 3 & 1 & & & & 0 \\ & 1 & \ddots & & & & \\ & & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \\ & 0 & & & & 3 & 1 \\ & & & & & 1 & 3 \end{pmatrix}.$$

Let d_n be $\det(A_n)$

Using cofactor expansion, we have:

$$d_n = 3d_{n-1} - d_{n-2}$$

EUCLIDIAN NRM DISTANCE:

The Euclidian norm of a vector $a \in \mathbb{R}^P$ is denoted as

$$||a||_2 = \sqrt{\sum_i^P a_i^2}$$

Euclidian distance between two vectors $a, b \in \mathbb{R}^P$

$$||a - b||_2 == \sqrt{\sum_i^P (a_i - b_i)^2}$$

DOT PRODUCT AND PROJECTION:

The dot product of two P-dimensional vectors a=

METRIC, MEASUREMENT, MEASURE:

METRIC: unit of measurement providing way to objectively quantify performance

MEASUREMENT: act of obtain data associated vs metric

MEASURES: numerical values associated with metric.

DATA CLASSIFICATION by measurement scales:

CATEGORICAL (NOMINAL) DATA - sorted into categories according to specified characteristics.

ORDINAL DATA - can be ordered or ranked according to some relationship to 1 another.

INTERVAL DATA - ordinal but have constant differences between observations; have arbitrary zero points.

RATIO DATA - continuous; have natural zero.

DATA RELIABILITY: VALIDITY:

(1st) RELIABILITY: Data is accurate; consistent

(2nd) VALIDITY: Data correctly measures what it is supposed to measure.

A tire pressure gage that consistently reads several pounds of pressure below true value	×	✓
Number of calls to customer service desk (counted correctly) used to assess customer dissatisfaction	✓	×
Customer rating on food quality is used to assess customer satisfaction	×	×

PERMUTATION: A permutation of set of objects is ordering of objects in row. $n \times (n-1) \times \ldots \times 1 = n!$

$r - permutations$ of set of n elements are:

$$P(n,r) = \frac{n!}{(n-r)!}$$

REMARK: $P(n,2) + P(n,1) = n^2$

COMBINATION: r − combination of set of n elements

$$\binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n!}{r!\,(n-r)!}$$

BINOMIAL COEFFICIENTS: For any $n \in \mathbb{Z}^+$, we have:

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y^1 + \cdots$$

$$= \sum_{i=0}^{n}\binom{n}{i}x^{n-i}y^i = v\sum_{i=0}^{n}\binom{n}{n-i}x^{n-i}y^i$$

NUMBER OF ELEMENTS IN POWER SET: $n \geq 0$, if set S has n elements, $N(\wp(S))$, total # subset of S

$$\sum_{K=0}^{n}\binom{n}{k} = \binom{n}{0} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n$$

$= N(\wp(S)) = 2^{|A|} = |P(S)|$

REMARK: $\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$

NUMBER OF INTEGER SOLUTIONS: # non-negative integer solutions of equation $x_1 + x_2 + \cdots + x_n = r$ OR # r-combinations with repetition allowed that can be selected from a set of n objects

$$\frac{(r+n-1)!}{r!\,(n-1)!} = \binom{n+r-1}{r}$$

ARRANGING IN A CIRCLE:

For n distinct objects arranged in a circle, there are

$$n!/n = (n-1)!$$

CONDITIONAL PROBABILITY of B given that A is

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(A)}}$$

$$\rightarrow \begin{pmatrix} \mathrm{P(A \cap B) = P(A)P(B|A)} \\ \mathrm{P(A \cap B) = P(B)P(A|B)} \end{pmatrix}$$

GENERAL MULTIPLICATION RULE:

$P(A_1A_2 \ldots A_n) =$

$P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\ldots$

$P(A_n|A_1A_2\ldots A_{n-1})$

INVERSE PROB:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

INDEPENDENT vs. MUTUALLY EXCLUSIVE:

Two events A & B being *independent*; *mutually exclusive* are *NOT same thing.*

A & B **independent** $\Leftrightarrow P(A \cap B) = P(A)P(B)$

A & B **mutually exclusive** $\Leftrightarrow P(A \cap B) = 0$

If A & B are *mutually exclusive & non-trivial (positive prob)* then A & B *cannot* be *independent.*

INDEPENDENCE VS. MUTUALLY EXCLUSIVE

Two events A & B being *independent*; *mutually exclusive* are *NOT same thing.*

A & B independent $\Leftrightarrow P(A \cap B) = P(A)P(B)$

A & B mutually exclusive $\Leftrightarrow P(A \cap B) = 0$

If A & B are *mutually exclusive & non-trivial (positive prob)* then A & B *cannot* be *independent.*

PAIRWISE INDEPENDENT EVENTS:

A set of events $A_1, A_2, \ldots, A_n$ are said to be pairwise independent $\Leftrightarrow P(A_iA_j) = P(A_i)P(A_j)$

MUTUALLY INDEPENDENT EVENTS:

A set of events $A_1, \ldots, A_n$ are said to be *mutually independent/ independent*

$$\Leftrightarrow P(A_1A_2\ldots A_k) = P(A_1)P(A_2)\ldots P(A_k)$$

$A_1, A_2, \ldots, A_n$ are mutually independent

$$\Leftrightarrow P(A_1A_2\ldots A_k) = P(A_1)P(A_2)\ldots P(A_k)$$

$Total\ 2^n - n - 1$ different cases.

Mutually independence → pair-wise independence

Pair-wise independence ↛ mutually independence

PARTITION: If $B_1, B_2, \ldots B_n$ are *mutually exclusive* $(B_iB_j = \emptyset;\ i \neq j)$; *exhaustive* $(B_1 \cup B_2 \cup \ldots \cup B_n = S)$ $\rightarrow B_1, B_2, \ldots, B_n$ a *partition* of S.

RULE OF TOTAL PROB.: If $B_1, B_2, \ldots, B_n$ is *partition*

$$P(A) = \sum_{i=1}^{n}P(B_iA) = \sum_{i=1}^{n}P(B_i)P(A|B_i)$$

$$= P(B_1)P(A|B_1) + \cdots + P(B_n)P(A|B_n)$$

BAYES'S THEOREM: Let $B_1, \ldots, B_n$ be partition of S.

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{P(B_1)P(A|B_1) + \cdots + P(B_n)P(A|B_n)}$$

$$= \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)} \rightarrow \frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \times \frac{P(A)}{P(A^C)}$$

CHEBYSHEV'S THEOREM: Proportion of values that lie within k $(k > 1)$ standard deviations of mean are at least $1 - 1/k^2$

$\mathrm{P}(|\mathrm{X} - \mu| \geq k\sigma) \leq \frac{1}{k^2} \rightarrow \mathrm{P}(|\mathrm{X} - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$

Why is this useful? Able to use mean; standard deviation to find percentage of total observations that fall within given interval about mean

DESCRIPTIVE ANALYSIS: characterise, consolidate; classify data to convert it into useful information for purposes of understanding; analysing business performance.

Measures of Location (Mean, Median, Mode)

Symmetrical, unimodal, mean = median = mode

Negatively skewed (left skewed, tails off toward right), mean < median < mode

Positive skewed (right skewed, tails off toward left), mode < median < mean

Measures of Dispersion (Range, Variance, Standard deviation, Chebyshev's Theorem, Coefficient of Variation)

Measures of Shape (Skewness, Kurtosis)

Measures of Association (Covariance; Correlation)

DESCRIPTIVE statistics for CATEGORICAL data:

PROPORTION is fraction of data that have certain characteristic, are key descriptive statistics for categorical data, i.e. defects or errors in quality control applications or consumer preferences in market research.

PREDICTIVE ANALYSIS:

—Seeks to predict future by examining historical data, detecting patterns or relationships in these data; then extrapolating relationships forward in time

—Predictive analysis can predict risks; find relationships in data not readily apparent with traditional analysis

—Using advantaged techniques, predictive analysis can help detect hidden pattern in large quantities of data to segment; group data into coherent sets to predict behaviour; detect trends

PRESCRIPTIVE ANALYSIS: optimise model (minimise expenditure, maximise benefit/profit)

Univariate statistics:

Discover associations between a variable of interest and potential predictors. It is strongly recommended to start with simple univariate methods before moving to complex multivariate predictors.

Most of univariate statistics are based on linear model which is one of main model in machine learning

RANDOM VECTORS & RANGE SPACE

Let E be an experiment; S a sample space. (X, Y) a two-dimensional random vector, range space is

$$R_{X,Y} = \{(x,y)|x = X(s), y = Y(s), s \in S\}$$

INDEPENDENT RANDOM VARIABLE:

X; Y are independent iff $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

$\Leftrightarrow$ X; Y are independent iff $f_{(X|Y)}(x|y) = f_X(x)$

$X_1, X_2, \ldots, X_n$ are independent iff

$f_{X_1,X_2,\ldots}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$

PERCENTILES: *kth percentile is value at or below which at least k percent of observations lie.*

COMPUTING PERCENTILES:

Find kth percentile for variable in sample size n

Rank of $kth\ percentile = nk/100 + 0.5$

QUANTILE: $q - th$ quantile of random variable X is z_q:

$$\mathrm{P}(\mathrm{X} \leq \mathrm{z_q}) = \mathrm{q} = \Phi(\mathrm{z_q}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mathrm{z_q}} \mathrm{e}^{-y^2/2}\mathrm{d}y$$

BREAK DATA INTO 4 PARTS

25th percentile, Q1; 50th percentile, Q2;

75th percentile, Q3; 100th percentile, Q4.

VARIANCE ~ average of squared deviations from mean. If sample data is also population data, then n = N to compute population variance

STANDARD DEVIATION ~ square root of variance (popular measure of risk)

STANDARD ERROR: $SE(X) = \sqrt{V(X)}/\sqrt{n}$

STANDARDIZED VALUES, Z-SCORE, provides relative measure of distance observation is from mean (independent of units of measurement)

$$\mathrm{z-score\ for}\ ith\ \mathrm{observation}\ z_i = \frac{x_i - \bar{x}}{s}$$

COEFFICIENT VARIATION provides relative measure of dispersion in data relative *to mean:*

CV = (standard deviation)/mean

Provides relative measure of risk to return

Useful when comparing variability of two or more data sets with different scales

Smaller CV → smaller risk

Reciprocal of CV → return to risk

COVARIANCE is measure of linear association between two variables, X; Y.

POSITIVE covariance → direct relationship

NEGATIVE covariance → inverse relationship

Magnitude → degree of association

❋$\sigma_{X,Y} = \mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$= E[(X - E(X))(Y - E(Y))]$

❋$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$

$= E(XY) - \mu_X\mu_Y$

❋$\mathrm{Cov}(X,X) = V(X)$

❋$\mathrm{Cov}(X,Y) = Cov(Y,X)$

❋$\mathrm{Cov}(aX + b, cY + d) = ac\ \mathrm{Cov}(X,Y)$

❋$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\ \mathrm{Cov}(X,Y)$

❋X, Y **independent**

$\rightarrow \mathrm{Cov(X,Y) = 0 \rightarrow E(XY) = E(X)E(Y)}$

CORRELATION is measure of linear association between two variables, X; Y (not dependent on units of measurement)

RANGE: −1 (Strong negative); 1 (Strong positive linear relationship); 0 indicates no linear relationship; Also known as: Pearson product moment correlation or Pearson's correlation coefficient

$$\mathrm{Cor}(X,Y) = \rho_{X,Y} = \rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

❋If X; Y are **independent**, then $\rho_{X,Y} = 0$.

On other hand, $\rho_{X,Y} = 0$ does not imply independence.

DISPERSION: Dispersion refers to degree of variation (numerical spread/ compactness)

RANGE = $max - min$ data values

INTERQUARTILE RANGE IQR difference between first; third quartiles: $(Q3 - Q1)$ (use 50% data)

SKEWNESS: describes lack of symmetry of data.

COEFFICIENT OF SKEWNESS (CS):

❋CS < 0 left-skewed data; CS > 0 right-skewed

❋|CS| > 1 suggests high degree of skewness.

❋0.5 ≤ |CS| ≤ 1 suggests moderate skewness.

❋|CS| < 0.5 suggests relative symmetry

KURTOSIS: refers to *peakedness* (high, narrow, or flatness)/(short, flat-top) of histogram

COEFFICIENT OF KURTOSIS CK: measures degree of kurtosis of population

CK <3 → data is flat and wide degree of dispersion

CK >3 → data is peaked with less dispersion

OUTLIERS: Mean; Range are sensitive to outliers

HOW DO WE IDENTIFY POTENTIAL OUTLIERS?

$$\mathrm{z-scores} > +3\ \mathrm{or} < -3$$

Extreme outliers are > 3 IQR to left Q1 or right Q3

Mild outliers are between 1.5 ∗ IQR; 3 ∗ IQR to left of Q1 or right of Q3

WHAT DO YOU DO WITH OUTLIERS?

❋Leave them in data if it is important

❋Remove them if they are different from rest

❋Correct error in data entry

DATA MODELLING; DIST. FITTING: Sample data limits our ability to predict uncertain events; potential values outside range of sample data are not included; better to identify underlying probability dist. from which sample data come by "fitting" theoretical dist. to data; verifying goodness of fit statistically

PROB. DIST.: characterization of possible values that random variable may assume along with probability of assuming these values

3 PERSPECTIVES FOR DEVELOPING

• theoretical arguments

• empirical data – empirical probability dist.

• using subjective values; expert judgement

Why do we need to know about distribution?

❋Helps you to understand underlying process that generates sample data. ❋Useful in building decision models with theoretical dist. of data.

❋Helps to compute probabilities of occurrence of outcomes to assess risk; make decisions

GOODNESS OF FIT: fitting data to probability dist.

CHI-SQUARE (need at least 50 data points)

KOLMOGOROV-SMIRNOV (works well for small samples; only for non-parametric data)

ANDERSON-DARLING (puts more weight on differences between tails of dist.)

SHAPIRO'S WILKINS NORMALITY TEST (test data against normal dist.) → P-value > 0.05 implies that dist. of data is not significantly different from normal dist. In other words, we can assume data is normal.

STANDARD NORMAL: X is called as standard normal random variable when $\mu = 0$; $\sigma = 1$; Z, $Z \sim N(0,1)$

Probability density $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

Distribution function $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy$

❋ $\mathrm{P(Z \geq 0) = P(Z \leq 0) = 0.5}$

❋ $\mathrm{Y \sim N(\mu, \sigma^2) \rightarrow X = \frac{Y - \mu}{\sigma} \sim N(0,1)}$

❋ $\mathrm{X \sim N(0,1) \rightarrow Y = aX + b \sim N(b, a^2)}$

❋ $\mathrm{X \sim N(\mu, \sigma^2)}$

$$\rightarrow \mathrm{P(a < X \leq b) = P}\left(\frac{\mathrm{a} - \mu}{\sigma} < \mathrm{Z} \leq \frac{\mathrm{b} - \mu}{\sigma}\right)$$

❋ $\mathrm{P(a < Z < b) = \Phi(b) - \Phi(a)}$

❋ $\mathrm{P(a < Z) = \Phi(a) \rightarrow\ P(a > Z) = 1 - \Phi(a)}$

❋ $\mathrm{P(|Z| < a) = 2\Phi(a) - 1}$

❋ $\mathrm{P(|Z| > a) = 2 - 2\Phi(a)}$

NORMAL DIST.: f(x) is bell-shaped curve

Properties: **1**. Symmetric; **2**. Mean = Median = Mode;

3. Range of X is unbounded;

4. Empirical rules apply (i.e., area under density function within ± 2 standard deviation is 95.4%, within ±3 standard deviation is 99.7%)

POISSON ≈ BINOMIAL: Let $X \sim B(n,p)$. $n \rightarrow \infty\ and\ p \rightarrow 0$; $\lambda = np$ remains a constant as $n \rightarrow \infty$.

$$\rightarrow X \sim P(np): \lim_{\substack{n \to \infty \\ p \to 0}} P(X = x) = \frac{e^{-np}(np)^x}{x!}$$

The approximation is good when $n \geq 20;\ p \leq 0.05$ OR $n \geq 100;\ np < 10$. If p is close to 1, we can still use Poisson distribution to approximate binomial probabilities.

NORMAL ≈ BINOMIAL

Use when: $n \rightarrow \infty\ and\ p \rightarrow 0;\ n \rightarrow \infty\ and\ p \rightarrow 0.5$

When n is small; p is not extremely close to 0 or 1, approximation is fairly good.

Use normal approximation only if $np > 5;\ n(1 - p) > 5$

CONTINUITY CORRECTION: X is binomial random variable mean $\mu = np, \sigma^2 = np(1-p) = npq$.

$$\text{Then as } n \rightarrow \infty: Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

❋$X \sim B(n,p): P(X = k) \approx P\left(k - \frac{1}{2} < X < k + \frac{1}{2}\right)$

❋$X \sim B(n,p): P(a \leq X \leq b) \approx$

$X \sim N(\mu, \sigma): P(a - 0.5 \leq X \leq b + 0.5)$

❋$X \sim B(n,p): P(a < X < b) \approx$

$X \sim N(\mu, \sigma): P(a + 0.5 < X < b - 0.5)$

GAMMA FUNCTION: (α is a complex number with positive real part). Gamma function $\Gamma(.)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy\ ;\ \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

❋ $\Gamma(1) = \int_0^\infty e^{-y}dy = 1$

❋ For integer, $\alpha = 1, 2, \ldots, n \rightarrow \Gamma(n) = (n-1)!$

χ^2 DISTRIBUTION: chi-square or χ^2 distribution with n degree of freedom is distribution of a sum of square of n independent standard random variables $\mathcal{N}(0,1)$.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z^2 \sim \chi^2(1) \rightarrow \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2(1)$

$$Y = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n) \rightarrow \sum_{i=1}^{n}Z_i^2 \sim \chi(n)$$

❋ For large n, $\chi^2(n) \sim N(n, 2n)$ approximately.

❋ If $(Y_1;\ Y_2;\ \ldots;\ Y_n)$ are independent chi-square random variables with $n_1, n_2, \ldots, n_k$ degree of freedom

$\rightarrow Y_1 + Y_2 + \cdots + Y_k$ has χ^2 distribution with $n_1 + \cdots + n_k$ degrees of freedom.

$$\sum_{i=1}^{k}Y_i \sim \chi^2\left(\sum_{i=1}^{k}n_i\right)$$

χ^2 – TABLE: χ^2 −table contains values of $\chi^2(n, \alpha)$ for various n: $P(Y \geq \chi^2(n; \alpha)) = \alpha; Y \sim \chi^2(n)$

$P(Y \geq \chi^2(n; 1 - \alpha)) = \alpha; Y \sim \chi^2(n)$

STUDENT t −DISTRIBUTION:

❋$n \rightarrow \infty \rightarrow \lim_{n \to \infty} f_T(t) = \frac{1}{\sqrt{2\pi}}\ e^{-z^2/2}$

❋$n \geq 30; X \sim t(n) \approx X \sim N(0,1)$.

❋The t-table *shows* $P(T > t_{n;\alpha}) = \alpha$

In table degree of freedom df = 10; $\alpha = 0.05$

$\rightarrow$ retrieve $t_{n;\alpha}$

If random sample was selected from a normal population

❋ $Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim N(0,1); U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

❋ $\bar{X}$ and S^2 are independent, so are Z and U

❋ $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}} \sim t(n-1)$

FISHERS F-DISTRIBUTION: *(ratio between two estimate of var.).* Random **samples** of size n_1 and n_2 are selected from 2 **normal population** with variances σ_1^2 and σ_2^2

$$U = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$

$$V = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

$$\rightarrow F = \frac{U/n_1}{V/n_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

❋$F \sim F(n,m) \rightarrow \frac{1}{F} \sim F(n,m)$

Table F-distribution gives value of $F(n_1, n_2, \alpha)$ such that

❋$P(F > F(n_1, n_2, \alpha)) = \alpha$ when $F \sim F(n_1, n_2)$

$$❋F(n_1, n_2; \alpha) = \frac{1}{F(n_1, n_2, 1 - \alpha)}$$

STATISTICAL SAMPLING

SAMPLING: foundation of statistical analysis.

ESTIMATORS: measures used to estimate unknown population parameters

POINT ESTIMATE: single number derived from sample that is used to estimate value of population parameters

UNBIASED ESTIMATOR: Let $\hat{\Theta}$ be estimator of θ(random var. based on sample). If $\mathrm{E}(\hat{\Theta}) = \theta$, $\hat{\Theta}$ is unbiased estimator of θ

$\bar{X}$ is an unbiased estimator of $\mu \rightarrow E(\bar{X}) = \mu$

An unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2\ ;\ E(S^2) = \sigma^2$$

A biased estimator of σ^2 is $T = \frac{1}{n}\sum(X_i - \bar{X})^2$

A SAMPLING EXPERIENCE: Sample size increases, average of sample means is all still close to expected value;

Standard deviation of sample means becomes smaller, meaning that means of samples are clustered closer together around true expected value.

SAMPLING DIST.: Sampling dist. of mean is dist. of means of all possible samples of fixed size n from some population. Standard deviation of sampling dist. of mean is called standard error of mean = $\sigma/\sqrt{n}$

❋As n increases, standard error↓, sampling error↓.

SAMPLING DISTRIBUTION related to SAMPLE MEAN:

$$\textbf{Sample mean}\ \bar{X} = \frac{1}{n}\sum X_i$$

Infinite population or from a finite population with replacement having mean μ_X; variance σ_X^2, sample distribution of sample mean $\bar{X}$ has mean; variance is:

❋$\boldsymbol{\mu_{\bar{X}}} = \mu_X$; $\boldsymbol{\sigma_{\bar{X}}^2} = \frac{\sigma_X^2}{n}$

❋$\boldsymbol{E(\bar{X})} = E(X)$

❋$\boldsymbol{V(\bar{X})} = V(X)/n$

❋$\boldsymbol{E(\bar{X} - \bar{Y})} = \mu_{\bar{X}-\bar{Y}} = \mu_1 - \mu_2$

❋$\boldsymbol{V(\bar{X} - \bar{Y})} = \sigma_{\bar{X}-\bar{Y}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

SAMPLING DISTRIBUTION relate to SAMPLE VARIANCE:

$$\textbf{Sample variance}\ S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

Let S^2 be sample variance of a random sample of size n taken from a normal population with $E(X) = \mu$; $V(X) = \sigma^2$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

$n - 1$ is degrees of freedom.

SAMPLING ERROR:

SAMPLING (STATISTICAL) ERROR: samples are only subset of total population

SAMPLING ERROR depends on size of sample relative to population.

NON-SAMPLING ERROR: sample does not adequately represent target population, results from poor sample design or choosing wrong population frame. (e.g., convenience sample)

EMPIRICAL RULES: For normally distributed data set, proportion of values that lie within k $(k > 1)$ standard deviations of mean follow empirical rules:

Application of Empirical Rule - Process Capability Index $(\mathrm{C_p})$ *is measure of how well manufacturing process can achieve specifications*

Using sample of output, measure dimension of interest; compute total variation using third empirical rule. $\mathrm{C_p}$ = (upper spec.−lower spec.)

total variation

ESTIMATING SAMPLING ERROR EMPIRICAL RULE

Using empirical rule for 3 standard deviations away from mean, ~99.7% of sample means should be between:

$[2.55, 7.45] -\ n = 10$; $[3.65, 6.35] -\ n = 25$

$[4.09, 5.91] -\ n = 100$; $[4.76, 5.24] - n = 500$

❋ n increases, standard error↓, sampling error↓.

LAW OF LARGE NUMBER LLN: Let $(X_1; X_2; \dots; X_n)$ be a random sample of size n with mean μ; variance σ^2. Then, $\forall \varepsilon \in \mathbb{R} \to (P(|\bar{X}-\mu| > \varepsilon) \to 0 \; as \; n \to \infty)$

CENTRAL LIMIT THEOREM:

Let $(X_1; X_2; \dots; X_n)$ be a random sample of size n with mean μ; variance σ^2.

$$Z = \frac{\bar X - \mu}{\sigma/\sqrt n} \sim N(0,1) \Leftrightarrow \bar X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\to P(\bar X > a) = P\left(Z > \frac{a-\mu}{\sigma/\sqrt n}\right)$$

Normal distribution provides an excellent approximation to sampling distribution of mean $\bar X$ if n ≥ 30.

❋If $(X_1; X_2; \dots; X_n)$ are (approximately) $N(\mu,\sigma^2)$, then $\bar X$ is (approximately) $N(\mu,\sigma^2/n)$ regardless of sample size n.

❋If sample size is large enough, then sampling dist. of mean ~normally distributed regardless of dist. of population ~ sample mean = population mean

❋If population ~ normally distributed, sampling dist. ~ normal distr. for any sample size.

INTERVAL ESTIMATES: 100(1 − a)% probability interval is any interval [A, B] such that probability of falling between; B is 1 − a. probability intervals are centred on mean/ median.

CONFIDENCE INTERVALS is range of values between which value of population parameter is believed to be, along with probability that interval correctly estimates true (unknown) population parameter.

$$P(\hat\Theta_L < \theta < \hat\Theta_U) = 1-\alpha$$

The interval computed $\hat\Theta_L < \theta < \hat\Theta_U$ is called $(1-\alpha)100\%$ confidence interval for θ. fraction $(1-\alpha)$ is called confidence coefficient or degree of confidence

❋For 95% confidence interval, if we chose 100 different samples, leading to 100 different interval estimates, we would expect that 95% of them would contain true population mean.

❋ Explain difference as level of confidence decreases from 95% to 90%.

When level of confidence decreases from 95% to 90%, range of CI increase → rejection area decreases.

CI for μ with KNOWN σ: z_α is number with an upper-tail probability of σ for standard normal distribution Z.

$$P(Z > z_\alpha) = \alpha \to \bar X \sim N\left(\mu,\frac{\sigma^2}{n}\right); Z = \frac{\bar X-\mu}{S/\sqrt n}\sim N(0,1)$$

We have $P\left(-z_{\alpha/2} \le \frac{\bar X-\mu}{\sigma/\sqrt n} \le z_{\alpha/2}\right) = 1-\alpha$

$$\to P\left(\bar X - z_{\alpha/2}\times\sigma/\sqrt n \le \mu \le \bar X + z_{\alpha/2}\times \sigma/\sqrt n\right) = 1-\alpha$$

SAMPLE SIZE FOR ESTIMATING μ: For margin of error e, confidence α, sample size $n \ge \left(\frac{z_{\alpha/2}\times\sigma}{e}\right)^2$

CONFIDENCE INTERVALS; SAMPLE SIZE

Determine appropriate sample size needed to estimate population parameter within specified level of precision (± E).

$\bar x \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt n}\right)$ and $E \ge z_{\alpha/2}\left(\frac{\sigma}{\sqrt n}\right)$

Sample size of mean: $n \ge (z_{\alpha/2})^2 \frac{\sigma^2}{E^2}$

Sample size for population:

$$\beta \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\beta(1-\beta)}{n}};$$

$$n \ge \left(z_{\frac{\alpha}{2}}\right)^2\frac{\pi(1-\pi)}{E^2}; \pi \text{ is proportion}$$

CONFIDENCE INTERVALS FOR SPECIAL CASES:

CI for PROPORTION: *Let $\hat p = x/n$ (sample proportion), where x is number in sample having desired characteristic; n is sample size.*

$$\hat p \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat p(1-\hat p)}{n}}$$

Suppose that we wish to determine number of voters to poll to ensure sampling error of at most ± 2%. With no information, use π = 0.5 (proportion who poll): $n \ge (1.96)^2(0.5)(1-0.5)/0.02^2$

Use sample proportion from preliminary sample as estimate of π or set π = 0.5 for conservative estimate to guarantee required precision (maximizes qty of π(1 − π)).

CI of μ; KNOWN σ with NORMAL population or n > 30	$\bar X \pm z_{\alpha/2}\frac{\sigma}{\sqrt n}$
CI of μ; UNKNOWN σ with NORMAL population & n < 30	$\bar X \pm t_{(n-1,\alpha/2)}\frac{S}{\sqrt n}$
CI of μ ; UNKNOWN σ with NORMAL population or n > 30	$\bar X \pm z_{\alpha/2}\frac{S}{\sqrt n}$
CI of $\mu_1-\mu_2$; KNOWN $\sigma_1^2 \ne \sigma_2^2$ NORMAL population or n > 30	$(\bar X_1-\bar X_2)\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
CI of $\mu_1-\mu_2$; UNKNOWN $\sigma_1^2 \ne \sigma_2^2$ with n > 30	$(\bar X_1-\bar X_2)\pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$

CI of $\mu_1-\mu_2$; UNKNOWN $\sigma_1^2 \ne \sigma_2^2$; NORMAL population; n < 30: Define

$$S_p^2 = \frac{\sum(X_i-\bar X)^2+\sum(Y_i-\bar Y)^2}{n_1+n_2+2} = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

is pooled sample variance, an estimator for $\sigma^2=\sigma_1^2=\sigma_2^2$

$100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is

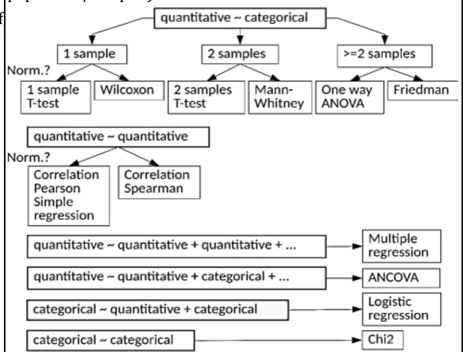
$$(\bar X_1-\bar X_2)\pm t_{n_1+n_2-2;\alpha/2}\,S_p\sqrt{\frac1{n_1}+\frac1{n_2}}$$

CI of $\mu_1-\mu_2$; KNOWN $\sigma_1^2 = \sigma_2^2$ with n_1; n_2 ≥ 30	$(\bar X_1-\bar X_2)\pm z_{\alpha/2}\,S_p\sqrt{\frac1{n_1}+\frac1{n_2}}$
CI of $\mu_D = \mu_1-\mu_2$ with NORMAL population, n < 30	$\bar D \pm t_{n-1;\alpha/2}\frac{S_D}{\sqrt n}$
CI of $\mu_D = \mu_1-\mu_2$ with n > 30	$\bar D \pm z_{\alpha/2}\frac{S_D}{\sqrt n}$
CI of σ^2; KNOWN μ with NORMAL population	$\left(\frac{\sum(X_i-\mu)^2}{\chi^2_{n;\alpha/2}};\frac{\sum(X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}\right)$
CI of σ^2; UNKNOWN μ with NORMAL population	$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}};\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right)$

CI of σ_1^2/σ_2^2; UNKNOWN μ_1; μ_2: NORMAL population

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\ \alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}$$

PAIRWISE ASSOCIATION TEST: On left is used non-parametric test of pairwise correlation (robust to non-normal population/samples)



Reference for statistical analysis using SAS, Stata, SPSS, R

https://stats.idre.ucla.edu/other/mult-pkg/whatstat/#

USING C.I. FOR DECISION MAKING:

1. Required volume for bottle-filling process is 800; sample mean is 796 mls. We obtained confidence interval for population mean of [790.12, 801.88]. Should machine adjustments be made? Although sample mean is less than 800, sample does not provide sufficient evidence to draw that conclusion that population mean is less than 800 because 800 is contained within confidence interval.

2. 1,300 voters found that 692 voted for particular candidate in two-person race. This represents proportion of 53.23% of sample. Could we conclude that candidate will likely win election? 95% confidence interval for proportion is [0.505, 0.559] This suggests that population proportion of voters who favour this candidate is highly likely to exceed 50%, so it is safe to predict winner.

3. What if sample proportion is 0.515; confidence interval for population proportion is [0.488, 0.543]? Even though sample proportion is larger than 50%, sampling error is large; confidence interval suggests that it is reasonably likely that true population proportion could be less than 50%, so you cannot predict winner.

PREDICTION INTERVALS is 1 that provides range for predicting value of new observation from same population.

While confidence interval is associated with sampling dist. of statistic, but prediction interval is associated with dist. of random variable itself.

$$\bar x \pm t_{\alpha/2,n-1}\left(s\sqrt{1+1/n}\right)$$

CONFIDENCE interval VS PREDICTION interval:

For 95% confidence interval means if we randomly choose n samples, there is 95% chance of them having desired values in ranges of 95% confidence level

For 95% prediction interval means that when there is new data coming, there is 95% chance that having desired values in 95% prediction level

HYPOTHESIS TESTING

Null hypothesis: What you do not want to see

Alternative hypothesis: what you want to see

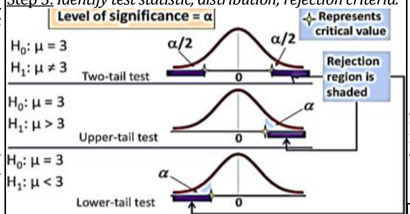
HYPOTHESIS TESTING PROCEDURE:

Step 1: Null; Alternative hypothesis (what you want to test). equal part is always in null hypothesis

Step 2: Determine level of significance (α); power (β).

	H_0 is true	*H_0 is false*
Reject H_0	Type I error P(reject $H_0 \mid H_0$ true) = α	Correct decision P(reject $H_0 \mid H_0$ false) =1 − β
Not reject H_0	Correct decision P(not reject $H_0 \mid H_0$ true) = 1 − α	Type II error P(not reject $H_0 \mid H_0$ false) = β

Step 3: *Identify test statistic, distribution; rejection criteria.*



Step 4: Compute test statistic value based on your data.

Step 5: Conclusion.

p − VALUE: If p − value > α, do not reject H_0, else reject H_0

IMPROVING POWER OF TEST

Power of test = 1 − b

• probability of not committing type II error

• should be high to make valid conclusion

How to ensure sufficient power?

❋Power of test is sensitive to sample size

❋small sample sizes → low power

❋large sample required for small α

HYPOTHESIS *on* μ: KNOWN σ: NORMAL population or $n > 30$:

To test: $\mu(>,<,=)\mu_0$

When H_0 is true, we have test statistic:

$$Z = \frac{\bar X-\mu_0}{\sigma/\sqrt n}\sim N(0,1)$$

HYPOTHESIS on μ: UNKNOWN σ NORMAL population

To test: $\mu(>,<,=)\mu_0$

When H_0 is true, we have test statistic:

$$T = \frac{\bar X-\mu_0}{S/\sqrt n}\sim t(n-1)$$

TWO-SIDED TEST ~ CONFIDENCE INTERVAL: $100(1-\alpha)\%$ confidence interval contains μ_0, $-t_{\alpha/2}\le t\le t_{\alpha/2}\to$ t is not located within rejection region → H_0 will NOT be rejected.

HYPOTHESIS on $\mu_1-\mu_2$ with KNOWN σ_1^2; σ_2^2: NORMAL population or $n_1, n_2 > 30$:

To test: $\mu_1-\mu_2(>,<,=)\delta_0$

When H_0 is true, we have test statistic:

$$Z = \frac{(\bar X_1-\bar X_2)-\delta_0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$$

HYPOTHESIS TEST ON σ_1^2, σ_2^2: To test: $\sigma_1^2 = \sigma_2^2$

We can use test statistic

$$F = \frac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$$

H_1	Rejection region
$\sigma_1^2 > \sigma_0^2$	$F > F_{n_1-1,n_2-1,\alpha}$
$\sigma_1^2 < \sigma_0^2$	$F < F_{n_1-1,n_2-1,1-\alpha}$
$\sigma_1^2 \ne \sigma_0^2$	$F < F_{n_1-1,n_2-1,1-\alpha/2}; F > F_{n_1-1,n_2-1,\alpha/2}$

HYPOTHESIS on $\mu_1-\mu_2$ with UNKNOWN σ_1^2; σ_2^2: $n_1, n_2 > 30$

To test: $\mu_1-\mu_2(>,<,=)\delta_0$

When H_0 is true, we have test statistic:

$$Z = \frac{(\bar X_1-\bar X_2)-\delta_0}{\sqrt{S_1^2/n_1+S_2^2/n_2}}\sim N(0,1)$$

HYPOTHESIS on $\mu_1-\mu_2$ with UNKNOWN $\sigma_1^2 = \sigma_2^2$: NORMAL population; $n_1, n_2 < 30$

To test: $\mu_1-\mu_2(>,<,=)\delta_0$

When H_0 is true, we have test statistic:

$$T = \frac{(\bar X_1-\bar X_2)-\delta_0}{S_p\sqrt{1/n_1+1/n_2}}\sim t(n_1+n_2-2)$$

$$S_p = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \text{ is}$$

pooled sample variance

HYPOTHESIS TEST ON PAIR SAMPLES:

For paired sample, define:

$$D_i = X_i - Y_i; \mu_D = \mu_1-\mu_2$$

To test: $\bar D(>,<,=)\mu_{D,0}$

When H_0 is true, we have test statistic:

❋$n < 30$, D_i are normally distributed

$$\to T = \frac{\bar D - \mu_{D,0}}{S_D/\sqrt n}\sim t(n-1)$$

❋$n \ge 30 \to Z = \frac{\bar D - \mu_{D,0}}{S_D/\sqrt n}\sim N(0,1)$

HYPOTHESIS TEST ON σ^2:

To test: $\sigma^2(>,<,=)\sigma_0^2$

We can use test statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$$

H_1	Rejection region
$\sigma_1^2 > \sigma_0^2$	$\chi^2 > \chi^2_{n-1,\alpha}$
$\sigma_1^2 < \sigma_0^2$	$\chi^2 < \chi^2_{n-1,1-\alpha}$
$\sigma_1^2 \ne \sigma_0^2$	$\chi^2 < \chi^2_{n-1,1-\alpha/2}$ or $\chi^2 > \chi^2_{n-1,\alpha/2}$

SAMPLE TEST ON PROPORTION:

$$z = \frac{(\hat p-\pi_0)}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Calculator with Normal distribution:

P(Z < 1.6) = MODE 3 → SHIFT 1 → 5 → 1 (P value) → 1.6

REJECTION REGION; P-VALUE for NORMAL distribution: $Z\sim N(0,1)$

H_1	*Rejection region*	*p-value*
>	$z > z_\alpha$	$P(Z > \lvert z\rvert)$
<	$z < -z_\alpha$	$P(Z < -\lvert z\rvert)$
≠	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z\rvert)$

t − distribution: $T\sim t(n)$

H_1	*Rejection region*	*p-value*
>	$t > t_{n,\alpha}$	$P(T > \lvert t\rvert)$
<	$t < -t_{n,\alpha}$	$P(T < -\lvert t\rvert)$
≠	$t > t_{n,\alpha/2}$ or $t < -t_{n,\alpha/2}$	$2P(T > \lvert t\rvert)$

T-TEST: Paired two-sample for means

TEST FOR EQUALITY OF VARIANCES between 2 samples using new type of test, F-test.

• To use this test, we must assume that both samples are drawn from normal populations.

• $H_0: \sigma_1^2-\sigma_2^2 = 0; H_1: \sigma_1^2-\sigma_2^2 \ne 0$

• F-test statistic: $F = s_1^2/s_2^2$

F-DIST. has two degrees of freedom, 1 associated with numerator of F-statistic, n_1-1; 1 associated with denominator , n_2-1.

Population with larger variance will be assigned numerator

ANALYSIS OF VARIANCE (ANOVA): Used to compare means of two or more population groups; fairly robust to departures from normality

$H_0: \mu_1 = \mu_2 = \dots = \mu_m$

H_1: at least 1 mean is different from others

• ANOVA measures variation between groups relative to variation within groups.

• Test statistic has F-dist. so if F-statistic is large enough based on level of significance chosen; exceeds critical value, we would reject H_0.

ANOVA Assumptions: Independence, Normality; homogeneity of variances:

1. Randomly; independently obtained (validated if random samples are chosen)

2. Normally distributed;

3. Have equal variances

If sample sizes are equal, violation of third assumption does not have serious effects, but with unequal sample sizes, it can.

Comparing sample means of two populations, use t-test rather than ANOVA

PEARSON CORRELATION TEST:

Test association between 2 quantitative variables:

The test calculates Pearson correlation coefficient; p-value for testing non-correlation. Let x; y be two quantitative variables, where n samples are observed. *linear regression coefficient is*

$$r = \frac{\sum(x_i-\bar x)(y_i-\bar y)}{\sqrt{\sum(x_i-\bar x)^2}\sqrt{\sum(y_i-\bar y)^2}}$$

Under H_0, test statistic $t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$ follows Student distribution with $n-2$ degree of freedom

NON−PARAMETRIC TEST OF PAIRWISE ASSOCIATION: When to use it? Observe data distribution: presence of outliers; distribution of residuals is not Gaussian.

Spearman rank-order correlation (quantitative ~ quantitative): measure of monotonicity of relationship between two datasets

Like other correlation coefficients, this one varies between −1; +1 with 0 implying no correlation.

Correlations of −1 or +1 imply an exact monotonic relationship.

Positive correlations imply that as $x\uparrow, y\uparrow$

Negative correlations imply that as $x\uparrow, y\downarrow$

Wilcoxon signed-rank test (quantitative ~ *cte*)

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It is equivalent to one-sample test of difference of paired samples.

It can be used as an alternative to paired Student's t-test, t-test for matched pairs, or t-test for dependent samples when population cannot be assumed to be normally distributed.

It has lower sensitivity compared to t-test. May be problematic to use when sample size is small

Null hypothesis H_0: difference between pairs follows a symmetric distribution around zero.

Mann- Whitney U test (quantitative ~ categorial 2 level): also called Mann-Whitney-Wilcoxon / Wilcoxon rank-sum test / Wilcoxon-Mann-Whitney test is a nonparametric test of null hypothesis that two samples come from same population against an alternative hypothesis, especially that a particular population tends to have larger values than other.

It can be applied on unknown distributions contrary to e.g. a t-test has to be applied only on normal distributions.

LINEAR MODEL: Given n random samples $(y_1, x_{1i}, \dots, x_{pi})$ linear regression models relation between observations y_i and independent variables x_i^p is $y_i = \beta_0+\beta_1x_{1i}+\dots+\beta_px_{pi}+\epsilon_i$

The $\beta's$ are regression coefficients

β_0 is intercept or bias

ϵ_i are residuals

Regression analysis is tool for building mathematical; statistical models that characterize relationships between dependent (ratio) variable; 1 or more independent, or explanatory variables (ratio or categorical), all of which are numerical.

Simple linear regression involves single independent variable >< multiple linear regression

Residuals are observed errors associated with estimating value of dependent variable using regression line: $e_i = Y_j-\hat Y_i$

Help detect outliers that bias regressions analysis.

Errors associated with individual observation

Residual analysis; Regression assumption

Residual = Actual Y value − Predicted Y value

Standard residual = residual/standard deviation

Rule of thumb: Standard residuals outside of ±2 *or* ± 3 are potential outliers.

An independent variable (*IV*), also called **predictors.** It is a variable that stands alone and isn't changed by other variables you are trying to measure. For example, someone's age might be an independent variable. Other factors (such as what they eat, how much they go to school, how much television they watch) aren't going to change a person's age. In fact, when you are looking for some kind of relationship between variables you are trying to see if independent variable causes some kind of change in other variables, or dependent variables.

A dependent variable, called a **target variable.** It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got night before you took test, or even how hungry you were when you took it. Usually when you are looking for a relationship between two things you are trying to find out what makes dependent variable change way it does.

Standard error − variability between observed; predicted Y values. This is formally called standard error of estimate, S_{YX}.

Step 1: model data: $y_i = \beta x_i + \beta_0 + \epsilon_i$
β: slope or coefficients or parameter of model
β_0: intercept or bias is second parameter of model
ϵ_i: $i-th$ error, or residual with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
Step 2: fit: estimate model parameters. goal is to estimate $\beta, \beta_0, and\ \sigma^2$
Minimises mean squared error MSE/Sum squared error SSE/Ordinary Least Squares OLS

$$\beta = \frac{1/n\ (\sum x_i y_i - \bar{y}\bar{x})}{1/n\ (\sum x_i - \bar{x})} = \frac{\text{Cov}(x, y)}{Var(x)}$$

FOUR MAJOR ASSUMPTIONS OF SIMPLE / MULTIPLE LINEAR REGRESSION:
1. Linearity (of relationship between Y & Xs): Residual vs. fitted - *Find straight horizontal line*
2. Normality of Errors = Errors (e; residuals) are normally distributed:
Normal Q-Q plot - Look for linear relationship
3. Homoscedasticity = Constant / Equal variance of errors (e) for all values of X = Impact of X on Y is same for all X values:
Residual vs. fitted; Scale-location - Look for straight horizontal line
4. Independence of errors = There is no correlation between errors (e) calculated from regression model - Need additional plot/test
* Residual time series plot
* Durbin-Watson test
• For cross-sectional data, this is usually not major issue
• Panel/time-series data need to check
• Issues 2, 3; 4 are often interrelated
• Cross-sectional data – data is collected only once, from different individuals/entities
• Panel/time-series data – data is collected multiple times from each individual/entity

MULTICOLLINEARITY: occurs when there are strong correlations among independent variables; they can predict each other better than dependent variable. Becomes difficult to isolate effect of 1 independent variable on dependent variable, signs of coefficients may be opposite of what they should be, making it difficult to interpret regression coefficients; p-values can be inflated. Correlations exceeding +0.7 may indicate multicollinearity

OVERFITTING: fitting model too closely to sample data at risk of not fitting it well to population in which we are interested.
R^2–value will increase if we fit higher order polynomial functions to data → make it difficult to explain phenomena rationally.
In multiple regression, if we add too many terms to model, then model may not adequately predict other values from population.
Overfitting can be mitigated by using good logic, intuition, theory; parsimony
❋Overfitting can be prevented by adding more data

PRINCIPLE OF PARSIMONY: Good models are as simple as possible

INTERACTIONS: occurs when effect of 1 variable is dependent on another variable. We can test for interactions by defining new variable as product of two variables, $X_3 = X_1 \times X_2$; testing whether this variable is significant, leading to alternative model.
Difference between correlation; interaction:
Whether two variables are associated says nothing about whether they interact in their effect on third variable. interaction between two variables means effect of 1 of those variables on third variable is not constant— effect differs at different values of other.

REGRESSION STATISTIC: Multiple R $-$ | r |, where r is sample correlation coefficient. r varies from -1 to +1 (r is negative if slope is negative).

F-TEST:
Goodness of fit: of a statistical model describes how well it fits a set of observations. Measures of goodness of for typically summarizes discrepancy between observed values and values under model in equation. We will consider explained variance also known as co-efficient of determination, denoted $\boldsymbol{R^2 (R - squared)}$

$\boldsymbol{R^2}$ **(R-squared)** is measure of "fit" of line to data. value of R^2 will be between 0%; 100%. A value of 1.0 indicates perfect fit; all data points would lie on line; larger value of R^2 better fit.
R^2 value ↑, order of polynomial ↑;
The total sum of squares SS_{tot} is sum of squares explained by regression, SS_{reg} plus sum of squares of residuals unexplained by regression, SS_{res}, also called SSE such that $SS_{tot} = SS_{reg} + SS_{res}$.
The mean of y: $\bar{y} = \frac{1}{n}\sum y_i$
The total sum of squares, also called total squared sum of deviations from mean y: $SS_{tot} = \sum(y_i - \bar{y})^2$
The regression sum of squares, also called explained sum of squares: $SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$
$where\ \hat{y}_i = \beta x_i + \beta_0$ is estimated value of $\hat{y}_i$ given a value of experience x_i
The sum of squares of residuals, also called residual sum of squares RSS is: $SS_{res} = \sum(y_i - \hat{y}_i)^2$
$\boldsymbol{R^2}$ is explained sum of squares of errors. It is variance by regression divided by total variance

$$R^2 = \frac{\text{explained SS}}{\text{total SS}} = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

Adjusted R-squared - adjusts R^2 for sample size; number of X variables.
Why use adjusted R Square? R-squared has additional problems that adjusted R-squared is designed to address.
Problem 1: Add predictor to model, R-squared increases, even if due to chance alone. It never decreases → model with more terms appear to have better fit simply because it has more terms.
Problem 2: If model has too many predictors; higher order polynomials, it begins to model random noise in data. This is known as overfitting model; it produces misleadingly high R-squared values; lessened ability to make predictions. adjusted R-squared increases only if new term improves model more than would be expected by chance. It decreases when predictor improves model by less than expected by chance. adjusted R-squared can be negative, but it's usually not, always $<$ R $-$ squared
Test: Let $\hat{\sigma}^2 = SS_{res}/(n-2)$ be an estimator of variance of ϵ. two in denominator stems from two estimated parameters: intercept and coefficient.
Unexplained variance: $\frac{SS_{res}}{\hat{\sigma}^2} \sim \chi^2_{n-2}$
Explained variance: $\frac{SS_{reg}}{\hat{\sigma}^2} \sim \chi^2_1$
The single degree of freedom comes from difference between $\frac{SS_{tot}}{\hat{\sigma}^2} \sim \chi^2_{n-1}$ and $\frac{SS_{res}}{\hat{\sigma}^2} \sim \chi^2_{n-2}$
The fisher statistics of ratio of two variances:

$$F = \frac{Explained\ var}{Unexplained\ var} = \frac{SS_{reg}/1}{\frac{SS_{res}}{(n-2)}} \sim F(1, n-2)$$

Using F-distribution, compute probability of observing a value greater than F under H_0: $P(x > F|H_0)$: survival function (1—Cumulative distribution function) at x of given F-distribution
Notice p-value (Significance F): When p-value is less than threshold (significance level), justifies rejection of null hypothesis. Null hypothesis is rejected when p < 0.05; not rejected when p > 0 .05.
Rejecting H_0 indicates X explains variation in Y

HOW TO FIND MODEL BEST FIT WITH DATA?
Method 1. USE R2 CHANGE (FOR NESTED MODELS, LINEAR REGRESSION): Find most parsimonious model by using R2; Rationale - Parsimonious model is preferred if it fits data (at least) equal to more complex model; Two models are considered as "nested" if one is constrained version of other
(1) $Y = b_0 + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + e$
(2) $Y\ =\ b0\ +\ b1 * X1\ +\ b2 * X2\ +\ e$
→ (2) nested in (1) because they are same if b3=0
→ (2) is more "parsimonious" than (1); estimate less no. of coefficients (= parameters)
We prefer Model (2) over Model (1) if R2 change between Models (2); (1) are not statistically significant (= simpler but equally well fit data)

How to compare nested models in r?
1) Fit each Model (1); Model (2) → lm(model1) / lm(model2)
2) Use F-test to test R2 change is statistically different → aov(model 1, model 2)
3) If test cannot reject H0 (= No R2 difference) choose (2) ; otherwise stay with (1)
Method 2. USE INFORMATION CRITERION (IC; FOR NON-NESTED MODELS): Find best-performing model by using information criterion; Rationale - Best-performing model is preferred, considering its complexity; fit to data; Commonly used information criterion measures for model selection
- Akaike Information Criterion (AIC) or its adjusted version (AICc)
- Baysian Information Criterion (BIC) → AIC/BIC are transformed values of a function of residuals; smaller is better
How to SELECT A MODEL by using information criterion in R?
1. Fit candidate models → e.g., $lm(model1)$ / $nls(model2)$ / …
2. For each model, calculate AIC or BIC → $AIC(model1)$ / $AIC(model2)$ / $AIC(\dots)$ or use BIC
3. Choose model having smallest value of AIC or BIC
IMPORTANT!!!
— ICs are statistical measures; assume one candidate model is (close to) "TRUE" model
— "True model" - a model that represents true, exact relationship between Y; X(s)
— In practice, you CANNOT check this assumption; usually ok for multiple regression/time series
— You need to choose model fit measures most suitable for your model(s); data!!!

INTERPRET REGRESSION ANALYSIS RESULT?
(e.g., $\boldsymbol{Y = \beta_0 + \beta_1 \times X + \epsilon}$)
(y) Intercept (also called constant) = Mean value of Y if value of all Xs = 0
Coefficient of X = impact of X on Y;
t-value = statistical test on each coefficient; constant (0 or not)
F-statistic = statistical test on all coefficients; constant (all 0 or not)
Pseudo R^2 (for nonlinear regression; also can be "adjusted"). Model fit in context of nonlinear models is usually defined in two ways
These are mathematically different from R^2 in linear models; therefore "Pseudo" R^2
1) Degrees of improvement from intercept-only model - McFadden's Pseudo R^2
2) Use same idea of linear R^2; variance of Y explained by X - Efron's Pseudo R^2 - Efron's Pseudo R^2 works ok for simple regression where linearity assumption is not severely violated
R2 - Measure of fit for linear regression model. proportion of variance in dependent variable explained by exploratory variable(s) = 1 – (Sum of squares of residuals;
SSR / SST(Sum of square total) = Pearson Correlation Coefficient R (Y, Predicted Y from model) ^{2} = "R" squared!
SSR = $\sum(Y - \text{predicted } Y)^2$ = Unexplained variance of Y

$$SSR = SS_{reg} = \sum(\hat{y}_i - \bar{y})^2$$

SST = $\sum$(Y – mean Y)^2 = Total variance of Y

$$SST = SS_{tot} = \sum(y_i - \bar{y})^2$$

Adjusted R^2 (for linear regression): R^2 tend to increase as number of predictors in model increases. Adjusted R^2 calculates "accurate R^2" by penalizing R^2 with number of predictors; sample size = 1 – [(1 - R^2) * (n – 1) / (n – p – 1)];
n = sample size, p = no. of predictors
Example: A study to predict number of home runs scored in a softball league with 32 teams of 9 players, based on different material used to make bat (alloy, composite, aluminium, hybrid); player's experience playing in league. ❋What is appropriate number of independent variables for regression model?
(4 – 1) materials + year experience = 3 variables.

MULTIPLE REGRESSION:
Multiple linear regression is most basic supervised learning algorithm.
Given a set of regression, we assume model that generates data involves only a linear combination of input variables.

$$y(x_i, \beta) = \beta^0 + \beta^1 x_i^1 + \dots + \beta^p x_i^p$$

$$y(x_i, \beta) = \beta_0 + \sum_{j=1}^{p-1} \beta_j + x_i^j$$

Extending each sample with an intercept $x_i := [i, x_i] \in R^{P+1}$ allows us to use a more general notation based on linear algebra and write it as a simple dot product: $y(x_i, \beta) = x_i^T \beta$
Where $\beta \in R^{P+1}$ is a vector of weights that defined $P+1$ parameters of model. From now we have P $regressors$ and intercept.
Minimise Mean Squared Error MSE loss:

$$MSE(\beta) = \frac{1}{N}\sum\left(y_i - y(x_i, \beta)\right)^2$$

$$= \frac{1}{N}\sum(y_i - x_i^T \beta)^2$$

$X = [x_0^T, x_1^T, \dots, x_N^T]$ be an $N \times P + 1$ metric of N samples of P inputs features with one column of one and let be $y = [y_1, y_2, \dots, y_n]$ be a vector of N targets.
mean squared error MSE loss is

$$MSE(\beta) = \frac{1}{N}\left\|y - X\beta\right\|_2^2$$

The β that minimise MSE can be found by:

$$\beta = (X^T X)^{-1} X^T y$$

MULTIPLE REGRESSION categorical independent variable/factor: ANALYSIS of COVARIANCE (ANCOVA)
Analysis of covariance (ANCOVA) is a linear model that blends ANOVA and linear regression. ANCOVA evaluates whether population means of a dependent variable (DV) are equal across levels of a categorical independent variable (IV) often called a treatment, while statistically controlling for effects of other quantitative or continuous variables that are not of primary interest, known as covariates (CV).
Regression as analysis of variance: ANOVA conducts F-test to determine whether variation in Y is due to varying levels of X.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ANOVA test for significance of regression:
H_0: population slope coefficient = 0
H_1: population slope coefficient ≠ 0
One way AN(C)OVA:
ANOVA: one categorical independent variable, one factor
ANCOVA: ANOVA with some co-variates.
Two way AN(C)OVA: $Ancova$ with two categorical independent variables, two factors.
MULTIPLE COMPARISONS:
Note that under null hypothesis distribution of p-values is uniform.
Statistical measures:
True Positive (TP) equivalent to a hit. test correctly concludes presence of an effect.
True Negative (TN). test correctly concludes absence of an effect.
False Positive (FP) equivalent to a false alarm, Type I error. test improperly concludes presence of an effect. Thresholding at $p - value <$ 0.05 leads to FP.
False Negative (FN) equivalent to a miss, Type II error. test improperly concludes absence of an effect.
***Bonferroni* correction for multiple comparisons**
The $Bonferroni$ correction is based on idea that if an experimenter is testing P hypotheses, then one way of maintaining familywise error rate (FWER) is to test each individual hypothesis at a statistical significance level of $1/P$ times desired maximum overall level.
So, if desired significance level for whole family of tests is α (usually 0.05), then $Bonferroni$ correction would test each individual hypothesis at a significance level of α/P. For example, if a trial is testing P = 8 hypotheses with a desired α = 0.05, then $Bonferroni$ correction would test each individual hypothesis at α = 0.05/8 = 0.00625.

The False discovery rate (FDR) correction for multiple comparisons
FDR-controlling procedures are designed to control expected proportion of rejected null hypotheses that were incorrect rejections ("false discoveries"). FDR-controlling procedures provide less stringent control of Type I errors compared to familywise error rate (FWER) controlling procedures (such as $Bonferroni$ correction), which control probability of at least one Type I error. Thus, FDR-controlling procedures have greater power, at cost of increased rates of Type I errors.
Brain volumes study
The study provides brain volumes of grey matter (gm), white matter (wm) and cerebrospinal fluid) (csf) of 808 anatomical MRI scans.
1. Fetch demographic data demo.csv and tissue volume data (gm.csv, wm.csv, csf.csv).
2. Merge tables.
3. Compute Total Intra-cranial (tiv) volume.
4. Compute tissue ratios: gm/tiv, wm/tiv.
5. Descriptive analysis per site in excel file.
6. Visualize site effect of gm ratio using violin plot:
$site\ \times\ gm$.
7. Visualize age effect of gm ratio using scatter plot:
$age\ \times\ gm$.
8. Linear model ($Ancova$):
$gm\_ratio \sim age\ +\ group\ +\ site$

Multivariate statistics:
Multivariate statistics include all statistical techniques for analysing samples made of two or more variables.
The data set ($N \times P\ matrix\ X$) is a collection of N independent samples column vectors $[x_1, \dots, x_i, \dots, x_N]$ of length P

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ \vdots \\ -\mathbf{x}_i^T- \\ \vdots \\ -\mathbf{x}_P^T- \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1P} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{iP} \\ \vdots & & \vdots & & \vdots \\ x_{N1} & \cdots & x_{Nj} & \cdots & x_{NP} \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1P} \\ \vdots & & \vdots \\ & \mathbf{X} & \\ \vdots & & \vdots \\ x_{N1} & \dots & x_{NP} \end{bmatrix}_{N \times P}$$

FORECASTING TECHNIQUES:
Qualitative; Judgmental techniques rely on experience; intuition.
Historical analogy approach obtains forecast through comparative analysis with prior situations; **Delphi method** questions anonymous panel of experts 2-3 times in order to reach convergence of opinion on forecasted variable; **Indicators** are measures that are believed to influence behaviour of variable we wish to forecast. Indicators are often combined quantitatively into **index**, single measure that weights multiple indicators, thus providing measure of overall expectation; **Leading indicators:** series of measure change before variable change; **Lagging indicators:** series of measures that follow change of variable.

STATICALLY FORECASTING MODELS:
Time Serie–*stream of historical data, daily*
Have components such as:
1. random behaviour; **2.** trend: is gradual upward or downward movement of time series; **3.** seasonal effects: is 1 that repeats at fixed intervals of time, typically year, month, …; **4.** cyclical effects: describe ups; downs over much longer time frame, i.e. several years
Stationary time series have only random behaviour.

MOVING AVERAGE MODEL: smoothing method based on idea of averaging random fluctuations in time series to identify underlying direction in which time series is changing. Simple moving average forecast for next period is computed as average of most recent k observations. Larger values of k result in smoother forecast models since extreme values have less impact

EXPONENTIAL SMOOTHING MODEL:
Simple: $\hat{y} = \alpha y_{x-1} + (1-\alpha)\hat{y}_{x-2}$
α is called smoothing factor/ coefficient/ constant.
Value of α dictates how much weight is given to most recent observed value versus last expected value;
$\alpha \in [0,1]$: regulates importance of most recent observations A_t with respect to smoothed mean of previous values;
$\alpha \approx 0$: *assign an almost constant weight to all past observations;*
$\alpha \approx 1$: *assign an almost constant weight to all recent observations.*
Double *Rewrite simple exponential smoothing:*
$F_{t+1} = a_t = \alpha A_t + (1-\alpha)F_t$
$\rightarrow F_{t+k} = a_t + b_t k$
$\rightarrow$ Level: $a_t = \alpha A_t + (1-\alpha)(a_{t-1} + b_{t-1})$
$\rightarrow$ Trend: $b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$
Predicted value $F_{(t+k)}$ is a function of last estimates of level a_t + linear trend $b_t * k$
$\beta \in [0,1]$: modulates importance of most recent value of trend; with respect to *smoothed* trend of previous periods.
$\beta \approx 0$: *assign most weight to trends in past*
$\beta \approx 1$: *most recently exhibited trend is pre-dominant.*

Regression-based forecasting for Time Series with Linear trend: Simple linear regression can be applied to forecasting using time as independent variable.

AUTOCORRELATION: When autocorrelation is present, successive observations are correlated with 1 another; for example, large observations tend to follow other large observations; small observations also tend to follow 1 another.
In such cases, other approaches, called AUTOREGRESSIVE MODELS, are more appropriate.

Forecasting time series with Seasonality: When time series exhibit seasonality, different techniques provide better forecasts than ones we have described:
Multiple regression models with categorical variables for seasonal components;
HOLT-WINTER MODEL, similar to exponential smoothing models in that smoothing constants are used to smooth out variations in level; seasonal patterns over time.

Holt-winter model for forecasting time series Seasonality; Trend:
HOLT-WINTERS ADDITIVE MODEL applies to time series with relatively stable seasonality:
$$\hat{y}_{T+\tau} = a_T + \tau b_T + s_T$$
a_T is smoothed estimate of level at time T
b_T is smoothed estimate of change in trend value at time T
s_T is smoothed estimate of appropriate seasonal component at T
HOLT-WINTERS MULTIPLICATIVE MODEL applies to time series whose amplitude increases or decreases over time; is $\hat{y}_{T+\tau} = (a_T + \tau b_T)s_T$

Regression forecasting with Causal variable: In many forecasting applications, other independent variables besides time, i.e. economic indexes or demographic factors, may influence time series. Explanatory/causal models, often called econometric models, seek to identify factors that explain statistically patterns observed in variable being forecast, usually with regression analysis

Practice of forecasting: Judgmental; qualitative methods are used for forecasting sales of product lines; broad company; industry forecasts. Simple time-series models are used for short; medium-range forecasts. Regression methods are typically used for long term forecasts.

DATA MINING:
PARALLEL CO-ORDINATES CHART consists of set of vertical axes, 1 for each variable selected. For each observation, line is drawn connecting vertical axes. point at which line crosses axis represents value for that variable.
SCATTERPLOT matrix combines several scatter charts into 1 panel, allowing user to visualize pairwise relationships between variables.
A VARIABLE PLOT plots matrix of histograms for variables selected.

DIRTY DATA: Real data sets that have missing values or errors, are called "dirty"; need to be "cleaned" before analysing them.
❋*Approaches for handling missing data.*
❋*Eliminate records that contain missing data*
❋*Estimate reasonable values for missing observations, i.e. mean or median value*
❋*Use data mining procedure to deal with them. XLMiner has capability to deal with missing data in Transform menu in Data Analysis group.*
❋*Try to understand whether missing data are simply random events or there is logical reason.*
❋*Eliminating sample data indiscriminately could result in misleading information; conclusions about data*

CLUSTER ANALYSIS, also called **DATA SEGMENTATION,** is collection of techniques that seek to group or segment collection of objects (observations or records) into subsets or clusters, such that those within each cluster are more closely related to 1 another than objects assigned to different clusters.
❋Objects within clusters should exhibit high amount of similarity, whereas those in different clusters will be dissimilar.

CLUSTER ANALYSIS METHODS:
HIERARCHICAL CLUSTERING, data are *not partitioned* into *particular cluster* in single step. Instead, series of partitions takes place, which may run from single cluster containing all objects to n clusters, each containing single object. *Hierarchical clustering may be represented by two-dimensional diagram known as dendrogram, which illustrates fusions or divisions made at each successive stage of analysis.*
AGGLOMERATIVE clustering methods proceed by series of fusions of n objects into groups.
DIVISIVE clustering methods separate n objects successively into finer groupings.

AGGLOMERATIVE CLUSTERING METHODS:
SINGLE LINKAGE CLUSTERING (NEAREST-NEIGHBOR): Distance between groups is defined as distance between closest pair of objects, where only pairs consisting of 1 object from each group are considered. At each stage, closest 2 clusters are merged
COMPLETE LINKAGE CLUSTERING: distance between groups is distance between most distant pair of objects, 1 from each group
AVERAGE LINKAGE CLUSTERING: Uses mean values for each variable to compute distance between clusters
WARD'S HIERARCHICAL CLUSTERING: Uses sum of squares criterion

CLASSIFICATION METHODS seek to classify categorical outcome into 1 of two or more categories based on various data attributes.
❋For each record in database, we have categorical variable of interest; number of additional predictor variables.
❋*For given set of predictor variables, we would like to assign best value of categorical variable.*

MEASURING CLASSIFICATION: Find probability of making misclassification error; summarize results in classification matrix, which shows number of cases that were classified either correctly or incorrectly.

USING TRAINING; VALIDATION DATA:
Data mining projects typically involve large volumes of data. data can be partitioned into:
▪ training data set – has known outcomes; is used to "teach" data-mining algorithm
▪ validation data set – used to fine-tune model
▪ test data set – tests accuracy of model

CLASSIFYING NEW DATA: after classification scheme is chose; best model is developed based on existing data, we use predictor variables as inputs to model to predict output

CLASSIFICATION TECHNIQUES/MODELS:
k-NEAREST NEIGHBOURS (K-NN) ALGORITHM
❋Finds records in database that have similar numerical values of set of predictor variables
❋Measure Euclidean distance between records in training data set. nearest neighbour to record in training data set is 1 that has smallest distance from it.
❋If $k = 1$, then $1 - NN$ rule classifies record in same category as its nearest neighbour.
❋$k - NN$ rule finds k-Nearest Neighbours in training data set to each record we want to classify; then assigns classification as classification of majority of k nearest neighbours.
❋Typically, various values of k are used; then results inspected to determine which is best.
HOW TO CHOOSE VALUE K?
Selecting value of K in K-nearest neighbour is most critical problem. Small value of K means that noise will have higher influence on result i.e., probability of overfitting is very high. Large value of K makes it computationally expensive; defeats basic idea behind KNN (that points that are near might have similar classes).
$\Longrightarrow$ Simple approach to select k is $k = n^{1/2}$.
To optimize results, we can use CROSS VALIDATION. We can test KNN algorithm with different values of K. Model which gives good accuracy can be considered to be optimal choice.

DISCRIMINANT ANALYSIS *is technique for classifying set of observations into predefined classes.*
Uses predefined classes based on set of linear discriminant functions of predictor variables
Based on training data set, technique constructs set of linear functions of predictors, known as discriminant functions:
$$L = b_1X_1 + b_2X_2 + \cdots + b_nX_n + C$$
b_i are discriminant coefficients (weights), X_i are input variables (predictors), c is constant (intercept)
MAXIMUM NUMBER OF FUNCTIONS = number of groups−1, or number of variables in analysis, whichever is smaller.
❋The weights of determining discriminant functions are computed by maximizing variance between groups relative to variance within groups.
For new observation, each of discriminant functions is evaluated; observation is assigned to class i if $i-th$ discriminant function has highest value.

LOGISTIC REGRESSION is variation of linear regression in which dependent variable is categorical. Estimates probability of belonging to category using regression on predictor variables
Seeks to predict probability that output variable will fall into category based on values of independent (predictor) variables.
This probability is used to classify observation into category. Generally used when dependent variable is binary—takes on two values, 0 or 1

Classification using logistic regression:
Estimate prob. p that observation belongs to category 1, $P(Y=1)$, and, consequently, probability $1-p$ that it belongs to category 0, $P(Y=0)$.
Then use $\underline{cutoff\ value}$, typically 0.5, with which to compare p; classify observation into 1 of two categories.
Dependent variable is called logit, which is natural logarithm of $p/(1-p)$ – called odds of belonging to category 1.
form of logistic regression model is
$$ln(p/(1-p)) = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k$$
logit function can be solved for p:
$$p = 1/\left(1 + \exp\left(-(\beta_0 + \beta_1X_1 + \cdots + \beta_kX_k)\right)\right)$$

ASSOCIATION RULE MINING, often called affinity analysis, seeks to uncover associations and/or correlation relationships in large data sets. Association rules identify attributes that occur together frequently in given data set.
Market basket analysis, for example, is used determine groups of items consumers tend to purchase together.
Association rules provide information in form of if then (antecedent consequent) statements

MEASURING STRENGTH OF ASSOCIATIONS:
SUPPORT for (association) rule is percentage (or number) of transactions that include all items both antecedent; consequent.
CONFIDENCE of (association) rule is ratio of number of transactions that include all items in consequent as well as antecedent (namely, support) to number of transactions that include all items in antecedent
$$\text{confidence} = P(\text{consequent}|\text{antecendent}) = \frac{P(\text{antecedence; consequent})}{P(\text{consequent})}$$
EXPECTED confidence is number of transactions that include consequent divided by total number of transactions.
LIFT is ratio of confidence to expected confidence. Higher lift ratio, stronger association rule; value greater than 1.0 is usually good minimum.
Example: supermarket database has 100,000 point-of-sale transactions; 2000 include both; B items; 5000 include C;; 800 include A, B; C
Association rule: "If; B are purchased, then C is also purchased."
Support = 800/100,000 = 0.008
Confidence = 800/2000 = 0.40
Expected confidence = 5000/100000 = 0.05
Lift = 0.40/0.05 = 8
The lift ratio indicates how much more likely we are to encounter event; B are purchased, as compared to entire population of transactions.

Cause; Effect modelling:
Correlation analysis can help us develop cause-and effect models that relate lagging; leading measures.
Lagging measures tell us what often external business results such has happened; are as profit, market share, or customer satisfaction.
Leading measures predict what will happen; are usually internal metrics i.e. employee satisfaction, productivity; turnover.

MONTE CARLO:
Monte Carlo simulation: is process of generating random values for uncertain inputs in model, computing output variables of interest; repeating this process for many trials to understand dist. of output.
Perform following steps:
1. Develop visual model
2. Determine probability dist. that describes uncertain inputs in model
3. Identify output variables you wish to predict
4. Set number of trials or repetitions for simulations
5. Run simulation
6. Interpret results

Market basket analysis, for example, is used determine groups of items consumers tend to purchase together.
Association rules provide information in form of if then (antecedent consequent) statements. In other situations, historical data are not available; we can draw upon properties of common prob. dist. to help choose representative dist. that has shape that would most reasonably represent analyst's understanding about uncertain variable.
Uniform or triangular dist. are often used in absence of data.

Sampling methods:
Monte Carlo sampling selects random variates independently over entire range of possible values of distribution. Monte Carlo sampling is more representative of reality; should be used if you are interested in evaluating model performance under various what-if scenarios.

Confidence interval for Mean:
Each time you run simulation, you will obtain slightly different results.
Confidence interval: $\bar{x} + z_{\alpha/2}(s/\sqrt{n})$
Because Monte Carlo simulation will generally have very large number of trials, we may use standard normal z value instead of t-dist. in confidence interval formula.

Flaws of averages: evaluation of model output using average value of input is not necessarily equal to average value of outputs when evaluated with each of input values.
In newsvendor example, quantity sold is limited to smaller of demand; purchase quantity, so even when demand exceeds purchase quantity, profit is limited.
Using average values in models can conceal risk.

Monte Carlo using simulation using Fitted Distribution:
Sampling from empirical data has some drawbacks.
Empirical data may not adequately represent true underlying population because of sampling error.
Using empirical dist. precludes sampling values outside range of actual data.

Steps for "Fitting" theoretical dist.; computing goodness of fit:
Choose suitable theoretical model:
For instance, normal or power law model. This task is informal; descriptive statistics like histogram; skewness indicator of observed data can be valuable hints;
Estimate model parameters:
Each theoretical model has parameters, for instance, mean; standard deviation for normal model. This task consists of estimating most likely model parameters for empirical dataset;
Determine significance level:
This tricky step establishes how good observed data match theoretical model with estimated parameters. If computed significance level is beyond pre-defined threshold, goodness-of-fit hypothesis is accepted, otherwise it is rejected

Estimate model parameters:
The maximum likelihood estimation method (MLE) is most popular method to estimate dist. parameters from empirical sample.
It finds model parameters that maximize likelihood of observed data with respect to theoretical model.

	No Seasonality	Seasonality
No trend	*Simple moving average or simple exponential smoothing*	*Holt-winter no-trend smoothing model or multiple regression*
Trend	*Double exponential smoothing*	*Holt-winter addictive or Holt-winter multiplicative model.*
	Population of size N	**Sample of n observations**
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{(\sum_{i=1}^{N}(x_i-\mu)^2)}{N}$	$s^2 = \frac{(\sum_{i=1}^{N}(x_i-\bar{x})^2)}{n-1}$
Co-var.	$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$	$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n-1}$
Co-relation	$\rho_{xy} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$	$r_{xy} = \frac{\mathrm{cov}(X,Y)}{s_x s_y}$

Errors metrics; Forecast Accuracy:

Mean absolute deviation: focus on mean value of errors $MAD = \frac{\sum_{t=1}^{n}\lvert A_t - F_t\rvert}{n}$	Mean square error / deviation: focus on variance of errors $MSE(MSD) = \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}$
For all metrics, smaller values → better data	
Root mean square error focus on standard deviation of errors $RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$	Mean absolute percentage error: cannot be used if time series contains 0 (division by 0) $MAPE = \frac{\sum_{t=1}^{n}\left\lvert\frac{A_t - F_t}{A_t}\right\rvert}{n} \times 100$

	DISCRETE RANDOM VARIABLE	CONTINUOUS RANDOM VARIABLE
Probability Mass/ Density Function	$f(x_i) = p(x_i) = p_i = P(X = x_i)$ $f(x_i) \geq 0\ \forall x_i$ *and* $\sum_{x_i} f(x_i) = 1;$ $P(X \in E) = \sum_{x_i \in E} f(x_i)$	$P(a < X \leq b) = \int_a^b f_X(x)\,dx, a < b \in \mathbb{R}$ $P(a < X < b) = P(a < X \leq b) = P(a \leq X < b)$ $= P(a \leq X \leq b) = \int_a^b f(x)\,dx;\ \int_{-\infty}^{\infty} f(x)\,dx = 1$
Cumulative Distribution Function	$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) = \sum_{t \leq x} P(X = t)$ $P(a \leq X \leq b) = P(X \leq b) - P(X < a)$ $= F(b) - F(a^-)$ If only *possible values are integers* and if *a and b are integers*, $\mathrm{P(a \leq X \leq b) = P(X = a, a+1, \dots, b)}$ $= \mathrm{F(b) - F(a-1)}$	$F(x) = \int_{-\infty}^{x} f(t)\,dt$ If derivative exists, we have $f(x) = \frac{d}{dx}F(x)$ $\to P(a \leq X \leq b) = P(a < X \leq b) = F(b) - F(a)$
Mean/ Expectation/ Expected values	$\mu_X = E(X) = \sum x f(x)\,; E(X^2) = \sum x^2 f(x)$ $E[g(X,Y)] = \sum_x \sum_y g(x,y) f_{X,Y}(x,y)$	$\mu_X = E(X) = \int x f(x)\,dx; E(X^2) = \int x^2 f(x)\,dx$ $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dy\,dx$
	❋ $E(a_0 + a_1 X_1 + \cdots) = a_0 + a_1 E(x_1) + \cdots$ ❋ *a; b constant,* $E(a + bX) = a + bE(X).$	❋ $b = 0 \to E(aX) = aE(X)$ ❋ $a = 1 \to E(X + b) = E(X) + b$
Expectation/ Mean of Function	$E[g(X)] = \sum g(x) f_X(x)$	$E[g(X)] = \int g(x) f_X(x)\,dx$
	❋ $g(x) = x^k \to E[g(X)] = E(x^k)$ is called k^{th} moment of X *and* $E(X^2)$ is called second moment	
Variance	$\sigma_X^2 = \mathrm{V(X) = E[(X - \mu_X)^2]} = \sum (x - \mu_X)^2 F_X(x)$	$\sigma_X^2 = \mathrm{V(X) = E[(X - \mu_X)^2]} = \int (x - \mu_X)^2 f_X(x)\,dx$
Joint Prob. Mass/ Density Function	$f_{X,Y}(x,y) = P(X = x, Y = y)$ $\sum_x \sum_y f_{X,Y}(x,y) = \sum_x \sum_y P(X = x, Y = y) = 1$	$P(a \leq X \leq b; c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dy\,dx$ $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$
Marginal Distribution	$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$ $= \sum_y f_{X,Y}(x,y)$ $\to f_Y(y) = P(Y = y) = \sum_x f_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$
Conditional Probability Mass Function	$\sum_x f_{(X\mid Y)}(x\mid y) = 1$ *and* $\sum_y f_{(Y\mid X)}(y\mid x) = 1$	$\int_{-\infty}^{\infty} f_{X\mid Y}(x\mid y)\,dx = 1\,;\ \int_{-\infty}^{\infty} f_{Y\mid X}(y\mid x)\,dy = 1$
	Conditional probability mass/density function of X: X and Y **independent** $\Leftrightarrow f_{X,Y} = f_X(x) f_Y(y)$ for some x and y *y is constant* $\to f_{(X\mid Y)}(x\mid y) \geq 0$	$f_{(X\mid Y)}(x\mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, if $f_Y(y) > 0, y \in Y$ $f_X(x) > 0 \to f_{X,Y}(x,y) = f_{(Y\mid X)}(y\mid x) f_X(x)$

Determine significance level:

Fit normal distribution, use Shapiro-Wilk test: If p-value is lower than threshold (usually fixed to 0.05) then normality hypothesis is rejected.

Fit arbitrary distribution, use Kolmogorov-Smirnov test: If p-value is lower than given threshold, goodness-of-fit hypothesis is rejected.

Cash-budget model is process of projecting; summarizing company's cash inflows; outflows expected during planning horizon.

Most cash budgets are based on sales forecasts. Because of inherent uncertainty in sales forecasts, Monte Carlo simulation is appropriate tool for modelling cash budgets.

LINEAR OPTIMISATION MODELS:

Building linear optimization models:

Step 1. Identify decision variables – unknown values that model seeks to determine.

Step 2. Identify objective function – quantity we seek to minimize or maximize.

Step 3. Identify all appropriate constraints – limitations, requirements, or other restrictions that are imposed on any solution, either from practical or technological considerations or by management policy.

Step 4. Write objective function; constraints as math expressions

Linear optimization model (often called **linear program/LP**) has 2 basic properties.

1. objective function; all constraints are linear functions of decision variables: This means that each function is simply sum of terms, each of which is some constant multiplied by decision variable.

2. All variables are continuous: This means that they may assume any real value (typically, nonnegative).

How simplex method works? simplex method evaluates impact of constraints in terms of their contribution to objective function for each variable. For simple case of only 1 constraint, optimal (maximum) solution is found by simply choosing variable with highest ratio of objective coefficient to constraint coefficient.

Example 3: *Crebo* Manufacturing produces 4 types of structural support fittings. Machining centres have capacity of 280,000 minutes per year. Gross margin/unit; machining:

Product	Plugs	Rails	Rivets	Clips
Gross margin/unit	0.3	1.3	0.75	1.2
Minute/unit	1	2.5	1.5	2

How many units of each product type should produce to maximize gross profit margin?

Objective: Maximize gross profit margin

$= 0.3\,X_1 + 1.3\,X_2 + 0.75X_3 + 1.2X_4$

Constraints: $X_1, X_2, X_3, X_4 \geq 0$

$1X_1 + 2.5\,X_2 + 1.5X_3 + 2X_4 \leq 280{,}000$

Clips have highest marginal profit per unit of resource consumed.

Maximum possible production of clips

= 280,000 minutes ÷ minutes/unit

= 280,000 ÷ 2 = 140,000

Profit for maximum production of clips

= gross margin/unit * max possible production

= $1.20 * 140,000 = $168,000

Outcomes:

Unique optimal solution: there is exactly 1 solution that will result in maximum (or minimum) objective.

Alternative (multiple) optimal solution: objective is maximized (or minimized) by more than 1 combination of decision variables, all of which have same objective function value.

Unbounded solution: objective can be increased or decreased without bound (i.e., to infinity for maximization problem or negative infinity for minimization problem)

Infeasibility: no feasible solution exists

Sensitivity analysis for Decision Variable:

Sensitivity Analysis allows us to understand how optimal objective value; optimal decision variables are affected by changes in objective function coefficients, impact of forced changes in certain decision variables, or impact of changes in constraint resource limitations or requirements.

Sensitivity Analysis applies to changes in only 1 of model parameters at time; all others are assumed to remain at their original values

Reduced Cost: How much objective function coefficient needs to be reduced for nonnegative variable that is zero in optimal solution to become positive.

If variable is positive in optimal solution, its reduced cost is zero. If objective coefficient of any 1 variable that has positive value in current solution changes but stays within range specified by Allowable Increase; Allowable Decrease, optimal decision variables will stay same; however, objective function value will change.

Sensitivity analysis for Constraints:

SHADOW PRICE - how much objective function will change as right-hand side of constraint is increased by 1. Whenever constraint has positive slack, shadow price is zero.

When constraint involves limited resource, shadow price represents economic value of having additional unit of that resource.

Using sensitivity analysis:

If change in objective function coefficient remains within Allowable Increase; Allowable Decrease ranges, then optimal values of decision variables will not change. However, you must recalculate value of objective function using new value of coefficient.

If change in objective function coefficient exceeds Allowable Increase or Allowable Decrease limits, then you must re-solve model to find new optimal values.

If change in right-hand side of constraint remains within Allowable Increase; Allowable Decrease ranges, then shadow price allows you to predict how objective function value will change → Multiply change in right-hand side (positive if increase, negative if decrease) by value of shadow price. However, you must re-solve model to find new values of decision variables.

If change in right-hand side of constraint exceeds Allowable Increase or Allowable Decrease limits, then you cannot predict how objective function value will change using shadow price → You must re-solve problem to find new solution.

INTEGER OPTIMIZATION:

Solving models vs. General Integer Variable:

Decision variables that we force to be integers are called general integer variables.

Algorithms for integer optimization models first solve LP relaxation (no integer restrictions imposed); gradually enforce integer restrictions using systematic searches.

Sensitivity analysis for Integer Optimization:

Because integer models are discontinuous by their very nature, sensitivity information cannot be generated in same manner as for linear models

To investigate changes in model parameters, it is necessary to re-solve model.

Example 1: A company makes 110-inch wide rolls of thin sheet metal; slices them in smaller rolls of 12, 15; 30 inches.

A cutting pattern is configuration of number of smaller rolls of each type that are cut from raw stock. Six different cutting patterns are used.

	Size of End Item			
Pattern	12in.	15in.	30in.	Scrap
1	0	7	0	5in.
2	0	1	3	5in.
3	1	0	3	8in.
4	9	0	0	2in.
5	2	1	2	11in.
6	7	1	0	11in.

Demands for coming week are 500 12-inch rolls, 715 15-inch rolls; 630 30- inch rolls.

Problem is to develop model that will determine how many 110-inch rolls to cut into each of six patterns to meet demand; minimize scrap.

Model development: Let x_i be number of 110-inch rolls to cut using pattern. x_i *needs* to be whole number (general integer variable) because each roll that is cut generated different number of end items. *The* only constraints are end-item demand, non-negativity; integer restriction

$Min\ 5x_1 + 5x_2 + 8x_2 + 2x_4 + 11x_5 + 11x_6$

$1x_3 + 9x_4 + 2x_5 + 7x_6 \geq 500\ (12\ in.rolls)$

$7x_1 + x_2 + x_5 + x_6 \geq 715\ (15\ in.rolls)$

$3x_2 + 3x_3 + 2x_5 \geq 630\ (30\ in.rolls)$

$x_i \in \mathbb{Z}$

Workforce scheduling model is practical, yet highly complex, problem in many businesses i.e. food service, hospitals; airlines.

Typically, huge number of possible schedules exist; customer demand varies by day of week; time of day, further complicating problem of assigning workers to time slots.

	PROBABILITY MASS/DENSITY FUNCTION	MEAN; VARIANCE
Discrete uniform distribution	$f_X(x) = P(X = x) = \begin{cases} 1/k & x = x_1, x_2, \dots, x_k \\ 0 & otherwise \end{cases}$ $\sigma^2 = V(X) = \frac{1}{k}\sum (x_i - \mu)^2$	$E(X) = \frac{1}{k}\sum x_i$ $\sigma^2 = \frac{1}{k}\sum x_i^2 - \mu^2$
Continuous uniform distribution	$f_X(x) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	$E(X) = (a+b)/2$ $V(X) = (b-a)^2/12$
Bernoulli trials	*Experiment with 2 outcomes ("success"; "failure")* $f_X(x) = P(X = x) = p^x(1-p)^{1-x}\ \forall x \in (0,1)$	$E(X) = p$ $V(X) = p(1-p)$
Binomial distribution	$n \in \mathbb{Z}^+;\ 0 < p < 1, X \sim B(n,p), x = 0,1 \dots n$ $f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$	$E(X) = np$ $V(X) = np(1-p)$
Negative binomial distribution	*#trials before obtain k successes;* $X \sim NB(k,p)$ $f_X(x) = P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$	$E(X) = k/p$ $V(X) = (1-p)k/p^2$
Geometric distribution	$0 < p < 1, X \sim Geom(p), x = 1, 2, 3, \dots$ $f_X(x) = P(X = x) = (1-p)^{x-1}p; q = 1 - p$ $P(X > n) = (1-p)^n = q^n$ <u>Memoryless property of Geometric:</u> $\mathrm{P(X > n + k \mid X > n) = P(X > k) = q^k, n, k \ge 1}$	*#required trials until first success is achieved* $E(X) = 1/p$ $V(X) = (1-p)/p^2$ $\mathrm{Geom(p) = NB(1, p)}$
	#failures until first success is achieved $Y = X - 1$ $P(Y = y) = (1-p)^y p$	$E(X) = (1-p)/p$ $V(X) = (1-p)/p^2$
Poisson random variable	*# success in fixed interval/period/region* $\lambda > 0$; $X \sim Poisson(\lambda); x = 0,1,2, \dots$ $f_X(x) = P(X = x) = e^{-\lambda}\lambda^x/x!$	$E(X) = \lambda$ $V(X) = \lambda$
Exponential distribution	$\lambda > 0; X \sim Exp(\lambda) \to f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$ $P(X > t) = e^{-\lambda t}$ <u>Memoryless property of Exponential distribution:</u> $P(X > s + t \mid X > s) = P(X > t) = e^{-\lambda t}$	$E(X) = 1/\lambda$ $V(X) = 1/\lambda^2$
Normal distribution	$\mu \in \mathbb{R}, \sigma > 0; \mathrm{X \sim N(\mu, \sigma^2)};\ -\infty < x < \infty$ $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$E(X) = \mu$ $V(X) = \sigma^2$
χ^2 distribution	*n degree of freedom;* $\Gamma(.)$ *gamma func.*; $\mathrm{Y \sim \chi^2(n)}$ $f_Y(y) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{\frac{n}{2}-1} e^{-y/2}\ \ (y > 0)$	$E(Y) = 2$ $V(Y) = 2n$
Student's t distribution	$Z \sim N(0,1); \mathrm{U \sim \chi^2(n)}$ $T = \frac{Z}{(\sqrt{U/n})} \sim t(n), -\infty < t < \infty$ $f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma(n/2)}\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}$	$E(T) = 0$ $V(T) = \frac{n}{n-2}; n > 2$
The F–distribution	$U \sim \chi^2(n_1)$ *and* $V \sim \chi^2(n_2)$ $F = \frac{U/n_1}{V/n_2};\ f_F(x) = \frac{n_1^{n_1/2} n_2^{n_2/2} \Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{x^{\frac{n_1}{2}-1}}{(n_1 x + n_2)^{\frac{n_1+n_2}{2}}}\right)$	

<u>PERMUTATION:</u> A permutation of set of objects is ordering of objects in row. $n \times (n-1) \times \dots \times 1 = n!$

$r - permutations$ of set of n elements are:

$$P(n,r) = \frac{n!}{(n-r)!}$$

<u>REMARK:</u> $P(n,2) + P(n,1) = n^2$

<u>COMBINATION:</u> $r -$ combination of set of n elements

$$\binom{n}{r} = \frac{P(n,r)}{r!} = \frac{n!}{r!\,(n-r)!}$$

<u>BINOMIAL COEFFICIENTS:</u> For any $n \in \mathbb{Z}^+$, we have:

$$(x+y)^n = x^n + \binom{n}{1} x^{n-1}y^1 + \cdots$$

$$= \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i = v \sum_{i=0}^{n} \binom{n}{n-i} x^{n-i} y^i$$

<u>NUMBER OF ELEMENTS IN POWER SET:</u> $n \ge 0$, if set S has n elements, $N(\wp(S))$, total # subsets of S is 2^n.

$$\sum_{K=0}^{n} \binom{n}{k} = \binom{n}{0} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n$$

<u>COVARIANCE:</u> $\sigma_{X,Y} = \mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$$= E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$$

❋$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y$

❋$\mathrm{Cov}(X,X) = V(X);\qquad \mathrm{Cov}(X,Y) = Cov(Y,X)$

❋$\mathrm{Cov}(aX + b, cY + d) = acCov(X,Y)$

❋$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$

❋X, Y **<u>independent</u>** $\to \mathrm{Cov(X, Y) = 0 \to E(XY) = E(X)E(Y)}$

<u>CORRELATION COEFFICIENTS:</u>

$$\mathrm{Cor}(X,Y) = \rho_{X,Y} = \rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

❋If X; Y are **<u>independent</u>**, then $\rho_{X,Y} = 0$.

On other hand, $\rho_{X,Y} = 0$ does not imply independence.

<u>STANDARD ERROR:</u> $SE(X) = \sqrt{V(X)}/\sqrt{n}$

<u>STANDARD NORMAL</u>: X is called as <u>standard normal</u> random variable when $\mu = 0; \sigma = 1$; Z, then $Z \sim N(0,1)$

<u>χ^2 DISTRIBUTION:</u> chi-square or χ^2 distribution with n degree of freedom is distribution of a sum of square of n independent standard random variables $\mathcal{N}(0,1)$.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z^2 \sim \chi^2(1) \to \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n) \to \sum_{i=1}^{n} Z_i^2 \sim \chi(n)$$

❋ For large n, $\chi^2(n) \sim N(n, 2n)$ approximately.

❋ If $(Y_1; Y_2; \dots; Y_n)$ are <u>independent chi-square random variables</u> with $n_1, n_2, \dots, n_k$ degree of freedom $\to Y_1 + Y_2 + \cdots + Y_k$ has χ^2 distribution with $n_1 + \cdots + n_k$ degrees of freedom.

$$\sum_{i=1}^{k} Y_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

<u>χ^2 – TABLE:</u> χ^2 –table contains values of $\chi^2(n,\alpha)$ for various n:

$P\big(Y \ge \chi^2(n;\alpha)\big) = \alpha; Y \sim \chi^2(n)$

$P\big(Y \ge \chi^2(n;1-\alpha)\big) = \alpha; Y \sim \chi^2(n)$

<u>SAMPLING DISTRIBUTION RELATE TO SAMPLE VARIANCE:</u>

CI of μ; KNOWN σ with NORMAL population or $n > 30$	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
CI of μ; UNKNOWN σ with NORMAL population & $n < 30$	$\bar{X} \pm t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}$
CI of μ ; UNKNOWN σ with NORMAL population or $n > 30$	$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$
CI of $\mu_1 - \mu_2$; KNOWN $\sigma_1^2 \ne \sigma_2^2$ NORMAL population or $n > 30$	$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
CI of $\mu_1 - \mu_2$; UNKNOWN $\sigma_1^2 \ne \sigma_2^2$ with $n > 30$	$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
CI of $\mu_1 - \mu_2$; UNKNOWN $\sigma_1^2 \ne \sigma_2^2$; NORMAL population; $n < 30$: Define	

	DISCRETE RANDOM VARIABLE	CONTINUOUS RANDOM VARIABLE
Probability Mass/ Density Function	$f(x_i) = p(x_i) = p_i = P(X = x_i)$ $f(x_i) \ge 0\ \forall x_i\ and \sum_{x_i} f(x_i) = 1; P(X \in E)$ $= \sum_{x_i \in E} f(x_i)$	$P(a < X \le b) = \int_a^b f_X(x)\,dx, a < b \in \mathbb{R}$ $P(a < X < b) = P(a < X \le b)$ $= P(a \le X < b)$ $= P(a \le X \le b) = \int_a^b f(x)\,dx; \int_{-\infty}^{\infty} f(x)\,dx$ $= 1$
Cumulative Distribution Function	$F(x) = P(X \le x) = \sum_{t \le x} f(t) = \sum_{t \le x} P(X = t)$ $P(a \le X \le b) = P(X \le b) - P(X < a)$ $= F(b) - F(a^-)$ If only <u>*possible values are integers*</u>; if <u>*a; b are integers*</u>, $\mathrm{P(a \le X \le b) = P(X = a, a + 1, \dots, b)}$ $\mathrm{= F(b) - F(a-1)}$	$F(x) = \int_{-\infty}^{\infty} f(t)\,dt$ If derivative exists, we have $f(x) = \frac{d}{dx}F(x)$ $\to P(a \le X \le b) = P(a < X \le b)$ $= F(b) - F(a)$
Mean/ Expectation / Expected values	$\mu_X = E(X) = \sum xf(x); E(X^2) = \sum x^2 f(x)$ $E[g(X,Y)] = \sum_x \sum_y g(x,y) f_{X,Y}(x,y)$	$\mu_X = E(X) = \int xf(x)\,dx; E(X^2)$ $= \int x^2 f(x)\,dx$ $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dy\,dx$
	❋$E(a_0 + a_1X_1 + \cdots) = a_0 + a_1E(x_1) + \cdots$	❋ $b = 0 \to E(aX) = aE(X)$

$= N(\wp(S)) = 2^{|A|} = |P(S)|$

REMARK: $\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$

NUMBER OF INTEGER SOLUTIONS: # non-negative integer solutions of equation $x_1 + x_2 + \dots + x_n = r$ OR # r-combinations with repetition allowed that can be selected from a set of n objects is

$$\frac{(r+n-1)!}{r!\,(n-1)!} = \binom{n+r-1}{r}$$

ARRANGING IN A CIRCLE:
For n distinct objects arranged in a circle, there are

$$n!/n = (n-1)!$$

CONDITIONAL PROBABILITY of B given that A is

$$P(B|A) = \frac{P(A\cap B)}{P(A)}$$
$$\rightarrow \begin{pmatrix} P(A\cap B) = P(A)P(B|A) \\ P(A\cap B) = P(B)P(A|B) \end{pmatrix}$$

GENERAL MULTIPLICATION RULE:
$P(A_1A_2\dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\dots P(A_n|A_1A_2\dots A_{n-1})$

INVERSE PROB:
$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

INDEPENDENT vs. MUTUALLY EXCLUSIVE:
Two events A & B being *independent*; *mutually exclusive* are *NOT* same thing.
A & B **independent** $\Leftrightarrow P(A\cap B) = P(A)P(B)$
A & B **mutually exclusive** $\Leftrightarrow P(A\cap B) = 0$
If A & B are *mutually exclusive & non-trivial (positive prob)* then A & B *cannot* be *independent.*

INDEPENDENCE VS. MUTUALLY EXCLUSIVE
Two events A & B being *independent*; *mutually exclusive* are *NOT* same thing.
A & B independent$\Leftrightarrow P(A\cap B) = P(A)P(B)$
A & B mutually exclusive$\Leftrightarrow P(A\cap B) = 0$
If A & B are *mutually exclusive & non-trivial (positive prob)* then A & B *cannot* be *independent.*

PAIRWISE INDEPENDENT EVENTS:
A set of events $A_1, A_2, \dots, A_n$ are said to be pairwise independent $\Leftrightarrow P(A_iA_j) = P(A_i)P(A_j)$

MUTUALLY INDEPENDENT EVENTS:
A set of events $A_1, \dots, A_n$ are said to be *mutually independent/ independent*
$\Leftrightarrow P(A_1A_2\dots A_k) = P(A_1)P(A_2)\dots P(A_k)$
$A_1, A_2, \dots, A_n$ are mutually independent
$\Leftrightarrow P(A_1A_2\dots A_k) = P(A_1)P(A_2)\dots P(A_k)$
$Total\ 2^n - n - 1$ different cases.
Mutually independence $\rightarrow$ pair-wise independence
Pair-wise independence $\nrightarrow$ mutually independence

PARTITION: If $B_1, B_2, \dots B_n$ are *mutually exclusive* $(B_iB_j = \emptyset;\ i \neq j)$; *exhaustive* $(B_1 \cup B_2 \cup \dots \cup B_n = S)$
$\rightarrow B_1, B_2, \dots, B_n$ a *partition* of S.

RULE OF TOTAL PROB.: If $B_1, B_2, \dots, B_n$ is *partition*

$$P(A) = \sum_{i=1}^{n} P(B_iA) = \sum_{i=1}^{n} P(B_i)P(A|B_i)$$
$$= P(B_1)P(A|B_1) + \dots + P(B_n)P(A|B_n)$$

BAYES'S THEOREM: Let $B_1, \dots, B_n$ be partition of S.

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{P(B_1)P(A|B_1) + \dots + P(B_n)P(A|B_n)}$$

Probability density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Distribution function $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}dy$
❋ $P(Z \geq 0) = P(Z \leq 0) = 0.5$
❋ $Y \sim N(\mu, \sigma^2) \rightarrow X = \frac{Y-\mu}{\sigma} \sim N(0,1)$
❋ $X \sim N(0,1) \rightarrow Y = aX + b \sim N(b, a^2);\ a, b \in \mathbb{R}$
❋ $X \sim N(\mu, \sigma^2)$
$$\rightarrow P(a < X \leq b) = P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right)$$
❋ $P(a < Z < b) = \Phi(b) - \Phi(a)$
❋ $P(a < Z) = \Phi(a) \rightarrow P(a > Z) = 1 - \Phi(a)$
❋ $P(|Z| < a) = 2\Phi(a) - 1$
❋ $P(|Z| > a) = 2 - 2\Phi(a)$

QUANTILE: $q-th$ quantile of random variable X is z_q:
$$P(X \leq z_q) = q = \Phi(z_q) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_q} e^{-y^2/2}dy$$

POISSON $\approx$ BINOMIAL: Let $X \sim B(n,p)$. $n \rightarrow \infty$ *and* $p \rightarrow 0$; $\lambda = np$ remains a constant as $n \rightarrow \infty$.
Then $X \sim P(np)$: $\lim\limits_{\substack{n\to\infty \\ p\to 0}} P(X = x) = \frac{e^{-np}(np)^x}{x!}$
The approximation is good when $n \geq 20;\ p \leq 0.05$ OR $n \geq 100;\ np < 10$. If p is close to 1, we can still use Poisson distribution to approximate binomial probabilities.

NORMAL $\approx$ BINOMIAL
Use when: $n \rightarrow \infty$ *and* $p \rightarrow 0$; $n \rightarrow \infty$ *and* $p \rightarrow 0.5$
When n is small; p is not extremely close to 0 or 1, approximation is fairly good.
Use normal approximation only if $np > 5;\ n(1-p) > 5$
Continuity correction: Suppose X is a binomial random variable mean $\mu = np$, variance $\sigma^2 = np(1-p) = npq$.
Then as $n \rightarrow \infty$: $Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$
❋ $X \sim B(n,p)$: $P(X = k) \approx P\left(k - \frac{1}{2} < X < k + \frac{1}{2}\right)$
❋ $X \sim B(n,p)$: $P(a \leq X \leq b) \approx X \sim N(\mu, \sigma)$: $P(a - 0.5 \leq X \leq b + 0.5)$
❋ $X \sim B(n,p)$: $P(a < X < b) \approx X \sim N(\mu, \sigma)$: $P(a + 0.5 < X < b - 0.5)$

UNBIASED ESTIMATOR: Let $\widehat{\Theta}$ be estimator of θ(random var. based on sample). If $E(\widehat{\Theta}) = \theta$, $\widehat{\Theta}$ is unbiased estimator of θ
$\bar{X}$ is an unbiased estimator of $\mu \rightarrow E(\bar{X}) = \mu$
An unbiased estimator of σ^2 is
$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2;\ E(S^2) = \sigma^2$$
A biased estimator of σ^2 is $T = \frac{1}{n}\sum(X_i - \bar{X})^2$

SAMPLING DISTRIBUTION RELATED TO SAMPLE MEAN:
$$\textbf{Sample mean}\ \bar{X} = \frac{1}{n}\sum X_i$$
Infinite population or from a finite population with replacement having mean μ_X; variance σ_X^2, sample distribution of sample mean $\bar{X}$ has mean; variance is:
❋ $\mu_{\bar{X}} = \mu_X$; $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$
❋ $E(\bar{X}) = E(X)$
❋ $V(\bar{X}) = V(X)/n$
❋ $E(\bar{X} - \bar{Y}) = \mu_{\bar{X}-\bar{Y}} = \mu_1 - \mu_2$
❋ $V(\bar{X} - \bar{Y}) = \sigma_{\bar{X}-\bar{Y}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

LAW OF LARGE NUMBER LLN: Let $(X_1; X_2; \dots; X_n)$ be a random sample of size n with mean μ; variance σ^2. Then, $\forall \varepsilon \in \mathbb{R} \rightarrow (P(|\bar{X} - \mu| > \varepsilon) \rightarrow 0\ as\ n \rightarrow \infty)$

CENTRAL LIMIT THEOREM: Let $(X_1; X_2; \dots; X_n)$ be a random sample of size n with mean μ; variance σ^2.

$$\textbf{Sample variance}\ S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$
Let S^2 be sample variance of a random sample of size n taken from a normal population with $E(X) = \mu$; $V(X) = \sigma^2$
$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$
$n-1$ is degrees of freedom.

STUDENT t – DISTRIBUTION:
❋ $n \rightarrow \infty \rightarrow \lim\limits_{n\to\infty} f_T(t) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$
❋ $n \geq 30$; $X \sim t(n) \approx X \sim N(0,1)$.
❋ The t-table *shows* $P(T > t_{n;\alpha}) = \alpha$
In table degree of freedom df $= 10$; $\alpha = 0.05 \rightarrow$ retrieve $t_{n;\alpha}$
If random sample was selected from a normal population then
❋ $Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \sim N(0,1)$; $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
❋ $\bar{X}$ *and* S^2 are independent, so are Z *and* U
❋ $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{Z}{\sqrt{U/(n-1)}} \sim t(n-1)$

FISHERS F-DISTRIBUTION: *(ratio between two estimate of var.)*
Random **samples** of size n_1 *and* n_2 are selected from 2 **normal population** with variances σ_1^2 *and* σ_2^2
$$U = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1), V = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$
$$\rightarrow F = \frac{U/n_1}{V/n_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$
❋ $F \sim F(n,m) \rightarrow \frac{1}{F} \sim F(n,m)$
Table F-distribution gives value of $F(n_1, n_2, \alpha)$ such that
❋ $P(F > F(n_1, n_2, \alpha)) = \alpha$ when $F \sim F(n_1, n_2)$
❋ $F(n_1, n_2; \alpha) = \frac{1}{F(n_1, n_2, 1-\alpha)}$

CONFIDENCE INTERVAL: $P(\widehat{\Theta}_L < \theta < \widehat{\Theta}_U) = 1 - \alpha$
The interval computed $\widehat{\Theta}_L < \theta < \widehat{\Theta}_U$ is called $(1-\alpha)100\%$ confidence interval for θ. fraction $(1-\alpha)$ is called confidence coefficient or degree of confidence

CI for μ with KNOWN σ: z_α is number with an upper-tail probability of σ for standard normal distribution Z.
$$P(Z > z_\alpha) = \alpha \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)\ or\ Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$$
$$\text{We have } P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$
$$\rightarrow P(\bar{X} - z_{\alpha/2} \times \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2} \times \sigma/\sqrt{n}) = 1 - \alpha$$

SAMPLE SIZE FOR ESTIMATING μ: For margin of error e,
confidence α, sample size is $n \geq \left(\frac{z_{\alpha/2} \times \sigma}{e}\right)^2$

CONFIDENCE INTERVALS; SAMPLE SIZE
Determine appropriate sample size needed to estimate population parameter within specified level of precision ($\pm$ E).
$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\ and\ E \geq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$Sample\ size\ of\ mean: n \geq (z_{\alpha/2})^2\frac{\sigma^2}{E^2}$$
Sample size for population:

$$S_p^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n_1 + n_2 + 2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$
is pooled sample variance, an estimator for $\sigma^2 = \sigma_1^2 = \sigma_2^2$
$100(1-\alpha)\%$ *confidence interval for* $\mu_1 - \mu_2$ *is*
$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2;\alpha/2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

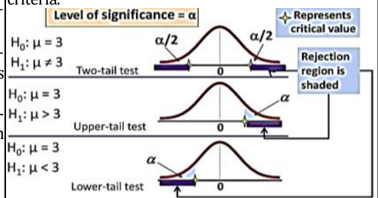
CI of $\mu_1 - \mu_2$; KNOW $\sigma_1^2 = \sigma_2^2$ with n_1 and $n_2 \geq 30$	$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
CI of $\mu_D = \mu_1 - \mu_2$ with NORMAL population, $n < 30$	$\bar{D} \pm t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}}$
CI of $\mu_D = \mu_1 - \mu_2$ with $n > 30$	$\bar{D} \pm z_{\alpha/2}\frac{S_D}{\sqrt{n}}$
CI of σ^2; KNOWN μ with NORMAL population	$\left(\frac{\sum(X_i - \mu)^2}{\chi^2_{n;\alpha/2}}; \frac{\sum(X_i - \mu)^2}{\chi^2_{n;1-\alpha/2}}\right)$
CI of σ^2; UNKNOWN μ with NORMAL population	$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right)$

CI of σ_1^2/σ_2^2; UNKNOWN μ_1; μ_2: NORMAL population
$$\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\ \alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}$$

HOW TO DO A HYPOTHESIS TEST
Step 1: Null; Alternative hypothesis (what you want to test)
The equal part is always in null hypothesis
Step 2: Determine level of significance (α); power (β).

	H_0 *is true*	H_0 *is false*
Reject H_0	Type I error $P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha$	Correct decision $P(\text{reject } H_0 \mid H_0 \text{ false}) = 1 - \beta$
Not reject H_0	Correct decision $P(\text{not reject } H_0 \mid H_0 \text{ true}) = 1 - \alpha$	Type II error $P(\text{not reject } H_0 \mid H_0 \text{ false}) = \beta$

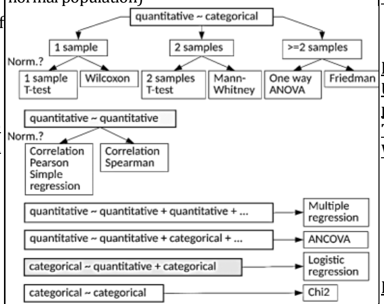
Step 3: Identify test statistic, distribution; rejection criteria.



Step 4: Compute test statistic value based on your data.
Step 5: Conclusion.

p – VALUE: If p – value $> \alpha$, do not reject H_0, else reject H_0

PAIRWISE ASSOCIATION TEST: On left is used non-parametric test of pairwise correlation (robust to non-normal population)



	❋ $a; b$ *constant*, $E(a + bX) = a + bE(X)$.	❋ $a = 1 \rightarrow E(X + b) = E(X) + b$
Expectation / Mean of Function	$E[g(X)] = \sum g(x)f_X(x)$	$E[g(X)] = \int g(x)f_X(x)\,dx$
	❋ $g(x) = x^k \rightarrow E[g(X)] = E(x^k)$ is called k^{th} moment of X *and* $E(X^2)$ is called second moment	
Variance	$\sigma_X^2 = V(X) = E[(X - \mu_X)^2] = \sum(x - \mu_X)^2F_X(x)$	$\sigma_X^2 = V(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty}(x - \mu_X)^2 f_X(x)\,dx$
	❋ $V(X) = E(X^2) - [E(X)]^2$; $V(X) = 0 \rightarrow P(X = \mu_X) = 1$; ❋ $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$	❋ $V(a + bX) = B^2V(X)$ ❋ $\sigma_X = SD(X) = \sqrt{V(X)}$
Joint Prob Mass/ Density Function	$f_{X,Y}(x,y) = P(X = x, Y = y)$ $\sum_x\sum_y f_{X,Y}(x,y) = \sum_x\sum_y P(X = x, Y = y) = 1$	$P(a \leq X \leq b; c \leq Y \leq d) = \int_a^b\int_c^d f_{X,Y}(x,y)dy\,dx$ $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\,dx = 1$
Marginal Distribution	$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x,y)$ $\rightarrow f_Y(y) = P(Y = y) = \sum_x f_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$
Conditional Prob Mass Function	$\sum_x f_{(X\|Y)}(x\|y) = 1$ *and* $\sum_y f_{(Y\|X)}(y\|x) = 1$	$\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)dx = 1$; $\int_{-\infty}^{\infty} f_{Y\|X}(y\|x)dy = 1$
	The conditional probability mass/density function of X: X; Y **independent** $\Leftrightarrow f_{X,Y} = f_X(x)f_Y(y)$ for some x; y y *is constant* $\rightarrow f_{(X\|Y)}(x\|y) \geq 0$	$f_{(X\|Y)}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, if $f_Y(y) > 0, y \in Y$ $f_X(x) > 0 \rightarrow f_{X,Y}(x,y) = f_{(Y\|X)}(y\|x)f_X(x)$

TWO-SIDED TEST ~ CONFIDENCE INTERVAL: $100(1-\alpha)\%$ confidence interval contains μ_0, $-t_{\alpha/2} \leq t \leq t_{\alpha/2} \rightarrow t$ is not located within rejection region $\rightarrow H_0$ will NOT be rejected.

HYPOTHESIS on $\mu_1 - \mu_2$ with KNOWN σ_1^2; σ_2^2: NORMAL population or $n_1, n_2 > 30$: To test: $\mu_1 - \mu_2 (>, <, =)\delta_0$
When H_0 is true, we have test statistic:
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

HYPOTHESIS TEST ON σ_1^2, σ_2^2: To test: $\sigma_1^2 = \sigma_2^2$
We can use test statistic
$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

H_1	Rejection region
$\sigma_1^2 > \sigma_0^2$	$F > F_{n_1-1,n_2-1,\alpha}$
$\sigma_1^2 < \sigma_0^2$	$F < F_{n_1-1,n_2-1,1-\alpha}$
$\sigma_1^2 \neq \sigma_0^2$	$F < F_{n_1-1,n_2-1,1-\alpha/2}$; $F > F_{n_1-1,n_2-1,\alpha/2}$

HYPOTHESIS on $\mu_1 - \mu_2$ with UNKNOWN σ_1^2; σ_2^2: $n_1, n_2 > 30$
To test: $\mu_1 - \mu_2 (>, <, =)\delta_0$
When H_0 is true, we have test statistic:
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim N(0,1)$$

HYPOTHESIS on $\mu_1 - \mu_2$ with UNKNOWN $\sigma_1^2 = \sigma_2^2$: NORMAL population; $n_1, n_2 < 30$
To test: $\mu_1 - \mu_2 (>, <, =)\delta_0$
When H_0 is true, we have test statistic:
$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$
$$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}\ \text{is}$$
pooled sample variance

HYPOTHESIS TEST ON PAIR SAMPLES:

	PROBABILITY MASS/DENSITY FUNCTION	MEAN; VARIANCE
Discrete uniform distribution	$f_X(x) = P(X = x) = \begin{cases} 1/k & x = x_1, x_2, \dots, x_k \\ 0 & otherwise \end{cases}$ $\sigma^2 = V(X) = \frac{1}{k}\sum(x_i - \mu)^2$	$E(X) = \frac{1}{k}\sum x_i$ $\sigma^2 = \frac{1}{k}\sum x_i^2 - \mu^2$
Continuous uniform distribution	$f_X(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$	$E(X) = (a+b)/2$ $V(X) = (b-a)^2/12$
Bernoulli trials	*Experiment with 2 outcomes ("success"; "failure")* $f_X(x) = P(X = x) = p^x(1-p)^{1-x}\ \forall x \in (0,1)$	$E(X) = p$ $V(X) = p(1-p)$
Binomial distribution	$n \in \mathbb{Z}^+$; $0 < p < 1, X \sim B(n,p), x = 0,1 \dots n$ $f_X(x) = P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$	$E(X) = np$ $V(X) = np(1-p)$
Negative binomial distribution	*#trials before obtain k successes;* $X \sim NB(k,p)$ $f_X(x) = P(X = x) = \binom{x-1}{k-1}p^k(1-p)^{x-k}$	$E(X) = k/p$ $V(X) = (1-p)k/p^2$
Geometric distribution	$0 < p < 1, X \sim Geom(p), x = 1, 2, 3, \dots$ $f_X(x) = P(X = x) = (1-p)^{x-1}p; q = 1 - p$ $P(X > n) = (1-p)^n = q^n$ Memoryless property of Geometric: $P(X > n + k \| X > n) = P(X > k) = q^k, n, k \geq 1$	*#required trials until first success is achieved* $E(X) = 1/p$ $V(X) = (1-p)/p^2$ Geom(p) = NB(1, p)
	#failures until first success is achieved $Y = X - 1$ $P(Y = y) = (1-p)^yp$	$E(X) = (1-p)/p$ $V(X) = (1-p)/p^2$
Poisson random variable	*# success in fixed interval/period/region* $\lambda > 0$; $X \sim Poisson(\lambda); x = 0,1,2, \dots$ $f_X(x) = P(X = x) = e^{-\lambda}\lambda^x/x!$	$E(X) = \lambda$ $V(X) = \lambda$
Exponential	$\lambda > 0; X \sim Exp(\lambda) \rightarrow f_X(x)$	$E(X) = 1/\lambda$ $V(X) = 1/\lambda^2$

$$= \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}$$

$$\rightarrow \frac{P(A|B)}{P(A^c|B)} = \frac{P(B|A)}{P(B|A^c)} \times \frac{P(A)}{(A^C)}$$

CHEBYSHEV'S INEQUALITY: Don't know how X behave

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \rightarrow P(|X-\mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

RANDOM VECTORS & RANGE SPACE

Let E be an experiment; S a sample space. (X, Y) a two-dimensional random vector. range space is

$$R_{X,Y} = \{(x,y) | x = X(s), y = Y(s), s \in S\}$$

INDEPENDENT RANDOM VARIABLE:

X; Y are independent iff $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

$\Leftrightarrow$ X; Y are independent iff $f_{(X|Y)}(x|y) = f_X(x)$

$X_1, X_2, \dots, X_n$ are independent iff $f_{X_1,X_2,\dots}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n)$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Leftrightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\rightarrow P(\bar{X} > a) = P\left(Z > \frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

Normal distribution provides an excellent approximation to sampling distribution of mean $\bar{X}$ if $n \geq 30$.

❋If $(X_1; X_2; \dots; X_n)$ are (approximately) $N(\mu, \sigma^2)$, then $\bar{X}$ is (approximately) $N(\mu, \sigma^2/n)$ regardless of sample size n.

GAMMA FUNCTION: (α is a complex number with positive real part). Gamma function $\Gamma(.)$ is defined by

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy; \ \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

❋ $\Gamma(1) = \int_0^{\infty} e^{-y} dy = 1$

❋ For integer values of $\alpha = 1,2,\dots,n \rightarrow \Gamma(n) = (n-1)!$

$$\beta \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\beta(1-\beta)}{n}}; \ n \geq \left(z_{\frac{\alpha}{2}}\right)^2 \frac{\pi(1-\pi)}{E^2}; \pi \text{ is proportion}$$

Suppose that we wish to determine number of voters to poll to ensure sampling error of at most ± 2%. With no information, use π = 0.5 (proportion who poll): $n \geq (1.96)^2 (0.5)(1-0.5)/0.02^2$

Use sample proportion from preliminary sample as estimate of π or set π = 0.5 for conservative estimate to guarantee required precision (maximizes qty of $\pi(1-\pi)$).

CONFIDENCE INTERVAL FOR PROPORTION

$\hat{p} = x/n$ (sample proportion), *where x is number in sample having desired characteristic; n sample size.*

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Reference for statistical analysis using SAS, Stata, SPSS, R

https://stats.idre.ucla.edu/other/mult-pkg/whatstat/#

HYPOTHESIS *on* μ+KNOWN σ:NORMAL population or n > 30:

To test: $\mu(>,<,=)\mu_0$

When H_0 is true, we have test statistic: $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

HYPOTHESIS on μ + UNKNOWN σ NORMAL population

To test: $\mu(>,<,=)\mu_0$

When H_0 is true, we have test statistic: $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$

For paired sample, define:

$$D_i = X_i - Y_i; \mu_D = \mu_1 - \mu_2$$

To test: $\bar{D}(>,<,=)\mu_{D,0}$

When H_0 is true, we have test statistic:

❋$n < 30, D_i$ are normally distributed

$$\rightarrow T = \frac{\bar{D} - \mu_{D,0}}{S_D/\sqrt{n}} \sim t(n-1)$$

❋$n \geq 30 \rightarrow Z = \frac{\bar{D} - \mu_{D,0}}{S_D/\sqrt{n}} \sim N(0,1)$

distribution	$P(X > t) = e^{-\lambda t}$ Memoryless property of Exponential distribution: $P(X > s + t \mid X > s) = P(X > t) = e^{-\lambda t}$	
Normal distribution	$\mu \in \mathbb{R}, \sigma > 0; X \sim N(\mu, \sigma^2); \ -\infty < x < \infty$ $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$E(X) = \mu$ $V(X) = \sigma^2$
χ^2 distribution	*n degree of freedom*; $\Gamma(.)$ *gamma func.*; $Y \sim \chi^2(n)$ $f_Y(y) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{\frac{n}{2}-1} e^{-y/2} \ (y > 0)$	$E(Y) = 2$ $V(Y) = 2n$
Student's t distribution	$Z \sim N(0,1); U \sim \chi^2(n)$ $T = \frac{Z}{(\sqrt{U/n})} \sim t(n), -\infty < t < \infty$ $f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{(n+1)/2}$	$E(T) = 0$ $V(T) = \frac{n}{n-2}; n > 2$
The F–distribution	$U \sim \chi^2(n_1)$ *and* $V \sim \chi^2(n_2)$ $F = \frac{U/n_1}{V/n_2}; \ f_F(x) = \frac{n_1^{n_1/2} n_2^{n_2/2} \Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{x^{\frac{n_1}{2}-1}}{(n_1 x + n_2)^{\frac{n_1+n_2}{2}}}\right)$	

HYPOTHESIS TEST ON σ^2: To test: $\sigma^2(>,<,=)\sigma_0^2$

We can use test statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

H_1	Rejection region
$\sigma_1^2 > \sigma_0^2$	$\chi^2 > \chi^2_{n-1,\alpha}$
$\sigma_1^2 < \sigma_0^2$	$\chi^2 < \chi^2_{n-1,1-\alpha}$
$\sigma_1^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{n-1,1-\alpha/2}$ or $\chi^2 > \chi^2_{n-1,\alpha/2}$

SAMPLE TEST ON PROPORTION:

$$z = \frac{(\hat{p} - \pi_0)}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Calculator with Normal distribution:

$P(Z < 1.6)$ = MODE 3 → SHIFT 1 → 5 → 1 (P value) → 1.6

REJECTION REGION; P-VALUE for NORMAL distribution: $Z \sim N(0,1)$

H_1	*Rejection region*	*p-value*
>	$z > z_\alpha$	$P(Z > \lvert z \rvert)$
<	$z < -z_\alpha$	$P(Z < -\lvert z \rvert)$
≠	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

***t* − distribution:** $T \sim t(n)$

H_1	*Rejection region*	*p-value*
>	$t > t_{n,\alpha}$	$P(T > \lvert t \rvert)$
<	$t < -t_{n,\alpha}$	$P(T < -\lvert t \rvert)$
≠	$t > t_{n,\alpha/2}$ or $t < -t_{n,\alpha/2}$	$2P(T > \lvert t \rvert)$

PEARSON CORRELATION TEST:

Test association between 2 quantitative variables:

The test calculates Pearson correlation coefficient; p-value for testing non-correlation

Let x; y be two quantitative variables, where n samples are observed. *linear regression coefficient is*

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

Under H_0, test statistic $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$ follows Student distribution with $n-2$ degree of freedom

NON - parametric test of PAIRWISE association:

When to use it? Observe data distribution: presence of outliers; distribution of residuals is not Gaussian.

Spearman rank-order correlation (quantitative ~ quantitative): measure of monotonicity of relationship between two datasets

Like other correlation coefficients, this one varies between -1; +1 with 0 implying no correlation.

Correlations of -1 or +1 imply an exact monotonic relationship.

Positive correlations imply that as $x \uparrow, y \uparrow$

Negative correlations imply that as $x \uparrow, y \downarrow$

Wilcoxon signed-rank test (quantitative ~ cte)

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It is equivalent to one-sample test of difference of paired samples.

It can be used as an alternative to paired Student's t-test, t-test for matched pairs, or t-test for dependent samples when population cannot be assumed to be normally distributed.

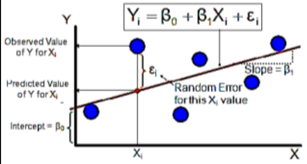
It has lower sensitivity compared to t-test. May be problematic to use when sample size is small

Null hypothesis H_0: difference between pairs follows a symmetric distribution around zero.

Mann- Whitney U test (quantitative ~ categorial 2 level): also called Mann-Whitney-Wilcoxon/Wilcoxon rank-sum test/Wilcoxon-Mann-Whitney test is a nonparametric test of null hypothesis that two samples come from same population against an alternative hypothesis, especially that a particular population tends to have larger values than other.

It can be applied on unknown distributions contrary to e.g. a t-test has to be applied only on normal distributions.

Linear model:



Given n random samples $(y_i, x_{1i}, \dots, x_{pi})$ with